148 P.

AMERIGAN MATHEMATICAL SOCIETY

Lecture Notes Prepared in Connection With the

Summer Seminar on Space Mathematics

held at

Cornell University, Ithaca, New York

July 1, 1963 - August 9, 1963

Part I

UNPUBLISHED PRELIMINARY DATA

Supported by the

National Aeronautics and Space Administration under Research Grant (NsG 358)

Air Force Office of Scientific Research under Grant AF-AFOSR 258-63

Army Research Office (Durham) under Contract DA-31-124-ARO(D)-82

Atomic Energy Commission under Contract AT(30-1)-3164

Office of Naval Research under Contract Nonr(G)00025-63

National Science Foundation under NSF Grant GE-2234.

RC
#1

TABLE OF CONTENTS

Part I

Summary of the Copenhagen Problem 1
Introduction to Dynamic Programming, by Dr. Richard E. Bellman 5
Calculus of Variations - Computational Aspects, by Dr. Richard E. Bellman 35
Dynamic Programming and Stochastic Control Processes, by Dr. Richard E. Bellman 47
Feedback Control Theory, Stochastic Control Theory, and Adaptive Processes, by Dr. Richard E. Bellman 69
Adaptive Control, by Dr. Richard E. Bellman 89
Elliptic Motion, by J.M.A. Danby 97
Matrix Methods, by J.M.A. Danby 140

Part II

Geodetic Problems and Satellite Orbits (Lectures I, II, III, IV, and V), by Dr. William H. Guier 151
Geodetic Problems and Satellite Orbits, by Dr. William H. Guier 219
Practical Astronomy, by Professor Paul Herget 255
Orbit Determination, by Professor Paul Herget 283
Calculus of Variations and Optimum Control Theory, by Magnus R. Hestenes 345
Rendezvous Problems, by J. C. Houbolt 409

Part III

Decay of Orbits, by P. J. Message 435
The Dominant Features of the Long Period Librations of the Trojan Minor Planets, by P. J. Message 455
Models of Gas Flows with Chemical and Radiative Effects, by F. K. Moore 467
The Stability Behavior of the Solutions of Hamiltonian Systems, by J. Moser 551
Lunar and Solar Perturbations on Artificial Satellites, by Peter Musen 595
Heat Transfer with Receding Boundaries and Other Complications, by Simon Ostrach 601

Part IV

Special Computation Procedures for Differential Equations, by S. V. Parter 701
Qualitative Methods in the n-Body Problem, by Professor H. Pollard 743
Outline of a Theory of Non-Periodic Motions in the Neighborhood of the Long-Period Librations about the Equilateral Points in the Restricted Problem of Three Bodies, by Professor E. Rabe 799
Elements of a Theory of Librational Motions in the Elliptical Restricted Problem, by Professor E. Rabe 823
Shock Waves in Rarefied Gases, by S. F. Shen 843
Basic Fluid Dynamics, by S. F. Shen 881
The Spheroidal Method in the Theory of Artificial Satellite Motion, by J. P. Vinti 927
Physical Experiments in Zero g Laboratories, by J. P. Vinti 939

Summary of the Copenhagen Problem - page 1

1	2	3	4	5	6	7	8	9	10	11	12	13
1	a	R	L_2	P	Sx		R	B B BT	3230 3251 3289		1894 1895	
2	b	R	L_3									See 1
3	c	R	L_1	P	Sx,y	L_1	∞	M	5374	60 99	1928 1935	
4	d	-	L_4	-	-	-	-	-	-	-	-	-
5	e	-	L_5	-	-	-	-	-	-	-	-	-
6	ℓ	R D*	m_1 m_2	P	Sx,y	∞	$L_{4,5}$	S	4015 4968	 30 47	1905 1918 1924	
7	-	R	-	P-A	Sx		—	S		61	1929	
8	m	R R*	m_1 m_2	P	Sx,y	∞	m_1 m_2	M	5286	48	1924	
9	k	D	m_1 m_2	P	Sx,y	m_1 m_2	$L_{4,5}$	S		80	1931	
10	f	R	m_2	P	Sx	m_2	∞	M		99	1935	
11	h	R	m_1									See 10
12	g	D	m_2	P	Sx	m_2	$L_{4,5}$?	S		39	1922	
13	i	D	m_1									See 12
14	n	R	-	P	Ay Sx		R	S		39 94	1922 1934	
15	r	R	-	P	Sx,y		$L_{4,5}$	M S	5374	60 97	1928 1934	
16	o	R	-	P	Ay Sx		$L_{4,5}$	M S	5374	60 97	1928 1934	
17	s	R	-	P	?		$L_{4,5}$	S		97	1934	
18	-	R	-	P-A	Ax Sy		—	S		64 67	1929 1930	

Explanation of symbols and list of columns.

1. Classification according to Strömgren's 1933 paper in the Bull. Astr. (2), 9, 87.

2. Classification according to Strömgren's 1925 paper in "Ergebnisse der exakten Naturwissenschaften, IV.

3. Motion relative to the rotating coordinate system

   R = retrograde
   D = direct

   Motion relative to the fixed coordinate system

   R* = retrograde
   D* = direct

4. Motion takes place around ...

5. Motion is periodic (P) or asymptotic (A).

6. Motion is symmetric (S) or asymmetrix (A) with respect to the rotating axes.

7. Class of orbits starts and ends, or reenters (R).
8.

9. Principal author

   S = Strömgren
   B = Burrau
   M = Möller
   T = Thiele

10. Astronomishe Nachrichten. No.

11. Copenhagen Obs. Publ. No.

12. Publication date.

13. Remarks.

Introduction to Dynamic Programming

by

Dr. Richard E. Bellman

Introduction to Dynamic Programming

In this series of five lectures, I would like to describe some of the fundamental ideas of dynamic programming, and some of its applications to the computational and analytic solution of problems in the calculus of variations, feed back control, trajectory optimization, and related problems. Many of these problems arise in a very natural way in connection with space travel. I would like first to turn to the purely mathematical aspects of the problem. Let us begin by discussing the classical approach of the calculus of variations, and point out some of its limitations. When we thus have motivation for finding some more efficient methods in certain cases, we will consider the approach of the methods of dynamic programming.

Suppose we start with a simple problem. We wish to minimize the functional

$$(1) \qquad \int_a^b g(u,u',t)dt$$

over all functions u(t) which, say, start out with a pre-assigned value. We have to find the optimal curve. This could be a trajectory which minimizes, for example, the time required to go between two points, or it might be the total amount of fuel consumed, or some combination of the two.

The problem is well set and straightforward. Classical calculus of variations says we obtain from the first variation of the integral an Euler equation,

$$(2) \qquad \frac{d}{dt}(g_{u'})-g_u=0$$

with associated initial condition u(a)=c and end condition

$$g_{u'}\Big|_{t=b}=0,$$

which is a necessary condition for a curve minimizing (1). At this point, most of the classical texts, if not all the classical texts, on the calculus of variations, close up shop and say the problem is solved. All that remains is to obtain a numerical solution. As a matter of fact, most often, they don't even mention that.

It's rather interesting to examine the philosophy of the concept of a numerical solution. Up until about the early 19th centruy, the mathematicians that existed would not have distinguished themselves very much from what we would now call physicists, applied mathematicians, or astronomers. To show the very close connection that existed between the two subjects, it is interesting to note that many held posts in what are still called in our country Departments of Mathematics and Astronomy. There was no question of the importance of a numerical solution. If Gauss or Newton was interested in a problem in celestial mechanics, he didn't feel that writing down an equation was a way of ending the problem. As far as they were concerned, this was a beginning of the problem.

Along about the beginning of the 19th century, with the great interest in rigorous foundations of mathematics, and rigorous derivations, a breed of mathematicians began to arise that was solely interested in the rigorous details and paid no attention to the applications. This split grew until at the present time not only do we have

mathematicians who do not work directly with numbers or with physical problems, but we have many hundreds of them who pride themselves in it and are very very proud of the fact that they wouldn't know what a resistor looked like if they stumbled over it, or what a nuclear reactor would look like even if a red light were flashing.

The separation between science and mathematics has been most unfortunate and, of course, it has had a most unfortunate influence upon mathematics, because very often one of the most interesting parts of the problem is the problem of actually getting a numerical solution to a numerical question. This certainly was one of the dicta of one of the greatest mathematicians of all time, Gauss. Until you had a feasible method for obtaining a numerical solution, you had no solution at all.

What has changed the picture very greatly, is we now have digital computers, which can do arithmetic very very fast. They can multiply a 10 digit number by another 10 digit number, in about a micro-second. Actually it is about six-tenths of a micro-second---let's say a 1/100,000 of a second---so they can do a hundred thousand multiplications of that type in a second. This means that there is a possibility that one can use methods which are quite different from the methods you might use if you only had desk computers, or slide rules.

One of the themes of my series of lectures will be to show that with the modern digital computer, we now have feasible methods available, which were not feasible before. Just as Poincaré said that a proof is a matter of the

fashion of the time, every proof is sufficient to the day thereof, so it is with computational solution. When you talk about the feasibility of a method, there's no absolute connotation. It's a function of what computational devices you have. If we had devices that could do arithmetic 10^{10} times faster than the devices we have now, we could use the crudest type of enumeration to solve some of the most complex and difficult problems around.

It is rather interesting to take a practical approach and find out what are some of the difficulties that you encounter. The first difficulty, you might say, is that Euler's equation (2) is only a necessary condition. Just as in ordinary calculus, if I want to find a minimum of a function over a given interval when I take the first variation and set it equal to zero, I might find a number of solutions, Fig. 1.

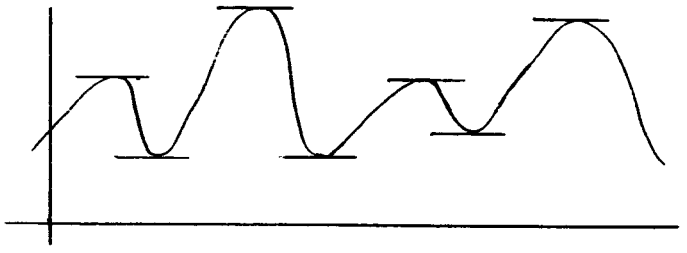


Fig. 1

How do I tell which one is the absolute minimum? Calculus gives us no method of doing that. Calculus says that I can give you necessary conditions for a local minimum or maximum, or sometimes of course, a point of inflection, something far more complicated, but I cannot give you

you any simple way of finding the absolute minimum. Now for a well-behaved function of a finite number of variables, you usually have only a finite number of critical points. Thus, in many cases, it's not too difficult to test for the absolute minimum.

For a problem in the calculus of variations, i.e. in infinite-dimensional space, it's easy to have a denumerable number of solutions satisfying these two-point boundary value problems. I think the easiest example of that is to take a torus, Fig. 2.

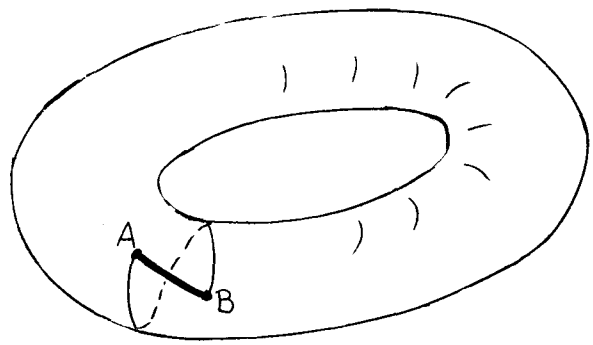


Fig. 2

Take two points A and B on the torus and say find the minimum distance between the two points, find the geodesic connecting A and B. Now, there is one curve (shown solid) connecting the two points which is the absolute minimum. Also, provided A and B are sufficiently close, there is a family of curves, one of which is shown, which winds around once. I can also find another family of curves which winds around twice and so on. In each one of these families there will be an absolute minimum for that family, and each one of these minima will be a relative minimum; and the solid line, the one that does not wind

around at all, will be the absolute minimum. This is a very nice way of showing that in a very sensible problem you can have a denumerable number of solutions.

Let us consider the computational aspects of the classical approach. Computationally, we face the problem of solving in general a nonlinear differential equation or nonlinear system subject to two-point conditions. Sometimes the two-point conditions come on because of the variational constraint. Sometimes they come on very naturally because we insist that we want to find the minimum time from one fixed point to another fixed point. If we look at the numerical problem involved and ask what can digital computers do if we're talking about large systems---5 dimensional, 10 dimensional, or 20 dimensional---digital computers can do one thing well. They can perform repetitive operations, which means that computers can solve initial value problems very very easily. The ideal problem for a digital computer is to solve ordinary differential equations subject to initial conditions. Unfortunately, in a calculus of variations, we have one initial condition missing at each point. The standard way to handle that is to, say, guess an initial derivative u'(a) and compute out families of curves, (Fig. 3), until we find one which fits the end condition at b.

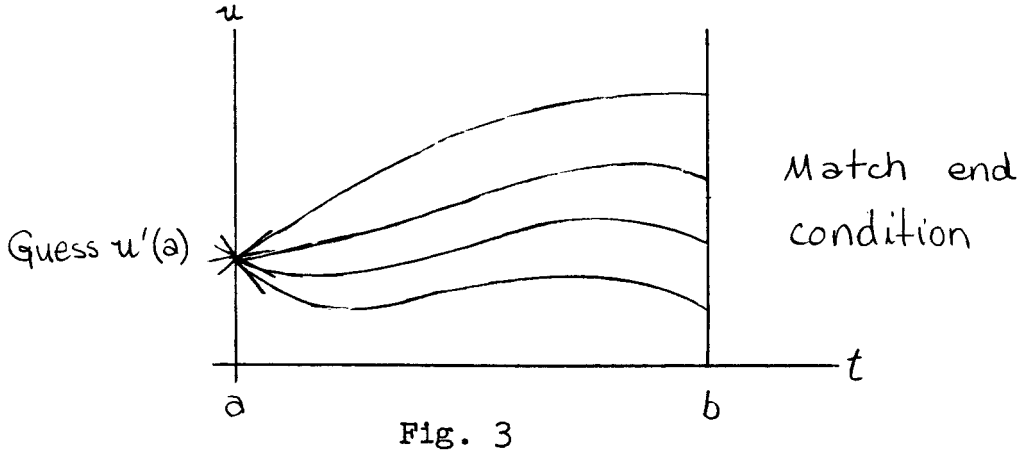


Fig. 3

We can use an interpolation method to zero in at t=b. This is the procedure which is used today with a certain amount of sophistication, but not much more sophistication.

There are several things which are wrong with this approach. In the first place, if this is a high-order equation, you're guessing points in, say, 4-dimensional, 8-dimensional, or 10-dimensional space. You might have to try a very large number of these trajectories. Secondly, many of these variational problems one can show are inherently unstable. This means, that a very small change in the initial conditions can produce a great change in the terminal conditions. Thirdly, (2) is only a necessary condition. One is finding a relative minima and, of course, since the Euler equation is just the equation of the first variation, one has to test that one is not also finding relative maxima or more complicated saddle point types of solutions by this trial and error approach. So, the two point boundary condition is a very important restriction on the applicability of the digital computer or any other type of computational technique.

Let's turn to more serious restrictions still. Suppose I introduce constraints on the optimizing function of the form, say,

$$|u'| \leqq k.$$

This could be interpreted as prohibiting optimal curves of the form

Fig. 4a

If we're talking about the motion of a rocket or an interceptor, or a missile, then it's clear that we don't want to consider motions like that. We don't concede any device that we have can do something like this.

Then, of course, you may have constraints on the function U itself. For example, you may say I want the altitude to be below a certain value, or above a certain value. I may have constraints of this type.

$$(3) \qquad a_1 \leqq u(t) \leqq b.$$

Actually, as far as the most important engineering and physical applications are concerned, the constraints are a very integral part of the program. Now, if one has constraints of the form (3) the situation becomes very complicated. Let's go back to the one-dimensional case. Suppose I have a function u(t) given over $[a,b]$.

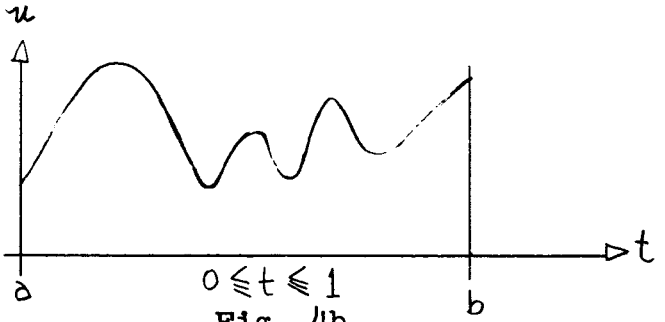


Fig. 4b

If I apply calculus, I get the turning points, but I know that if I have constraints I have to test the end points. Of course, this is deliberately drawn so that in this case the end points are the absolute minimum and the absolute maximum. Now what does this mean in the calculus

of variations? It means this. Suppose I draw u as a function of t.

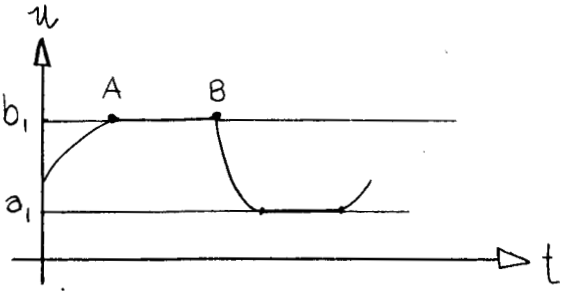


Fig. 5

What can happen (and it's easy to construct very simple examples in which this does happen) is that the solution is constructed in the following way: Starting from t=0, you follow an Euler trajectory until you hit a boundary. Then you go along the boundary for awhile; then you follow an Euler trajectory until you hit another boundary, and so on. You can see what the computational difficulties are. The points A, B, C, D, ..., where one hits the boundary are unknown. As a matter of fact, there is no simple way of determining how many different pieces the solution will have. There is no simple way of looking at a problem with constraints and determining, for example, whether it will hit the boundary at all! If it hits the boundary, does it stay on the boundary all the way or come off, and so forth? So if we take the computational difficulties that we had before, without constraints, and add constraints, you see that you have a very formidable problem.

There are classical techniques in the calculus of variations for handling constraints. What these do is introduce additional functions satisfying additional two-point boundary conditions. In some cases, if the constraints

are simple enough, one can apply classical techniques. A nice example of an important constraint is where one wants the derivative to be ± 1. For example, in the optimization of rockets, if you're talking about when you go full speed and when not---since there's so little control over the rocket engine---essentially all you can say is make the rate at which you burn fuel either +1 or 0; either burn at maximum rate or do not burn at all. Of course, that's a fairly simple example, because, in that case, one can show that there are two regions or three regions for which you have to pick a point at which you want to go at maximum rate, and so forth. But when I talk about control problems, I'll come back to this again.

This type of control, where there are only two values has the picturesque name, "Bang-Bang Control". It's very important from the engineering point of view, because it's clear that it's much easier to have a device which is either on or off than something which has to measure certain state variables and adjust itself to those variables. So this is a highly desirable type of control, which is why it's become of importance.

Now, suppose I introduce uncertainties, first in the form of stochastic elements. Then I can make the problem even more complicated and add adaptive elements. As an example, suppose we're trying to fly through an atmosphere which has certain unknown properties, and we have both to direct our course and determine something of the properties of the atmosphere at the same time. That would be an adaptive, or learning, process. I'll talk about these in more detail.

Let's just assume there are certain unknown features present---winds, small deviations of the atmosphere, small deviations in the way in which the engine runs, small errors in direction and, in order to get around the fact that we really don't know about these things, we assume that they're random variables. This is always an assumption, and what we'd like to do is replace these stochastic elements by adaptive elements. We hope that we can learn more about the unknown elements as we go along.

Now, of course, if we consider these more modern features, then the classical techniques are very difficult. This is some of the motivation for a reexamination of variational problems to see if we cannot tackle them by different methods which provide, in some cases, a better analytic formulation and/or a better computational approach. But I want to mention right now that in no way does a new approach supercede the application of an old approach. Generally speaking, it's very hard in mathematics to find any situation where one method completely replaces another. To a great extent, as new methods occur they complement the older methods, and it's the combination of the two together that is the most powerful. What we'll find is that in dynamic programming we're constructing a theory which is dual to the classical theory---dual in the classical geometric sense. I'll point this out again later. This makes it very clear that the two theories taken together will be very much more powerful than either one by itself.

Before discussing variational problems, let us take

a very simple problem in calculus in order to illustrate the approach we're going to use. Suppose I take the schoolboy calculus problem: A stone is thrown straight up with velocity V. What is the maximum height it attains? We know how to solve the problem in terms of calculus. We say let $x(t)$ represent the altitude of the stone at time t. We are assuming implicitly, of course, that once the stone is thrown up, it's acted upon only by gravity, which means that the acceleration is -g, directed downwards. The initial condition is that at time zero the altitude was zero, and we said the initial velocity was V; so we're as happy as could be, because we have a second order differential equation, $x''=-g$ with two initial values, $x(0)=0$ and $x'(0)=V$, which we can, if we insist upon it, solve computationally, and just trace out trajectories. In this case, of course, we can actually carry out the solution to obtain

(4) $\quad x'=V+gt, \quad x=Vt+gt^2/2.$

The trajectory over time is

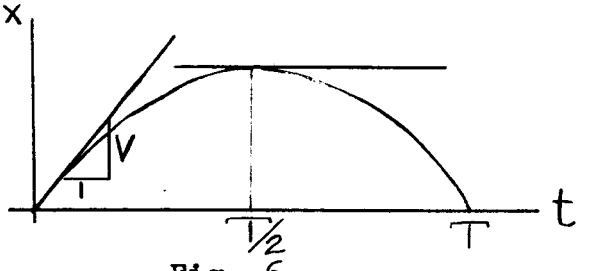


Fig. 6

which, we find, has a maximum value. To find the maximum value I differentiate $x(t)$ with respect to time and find where $x'=0$. Thus the time at which maximum altitude

is obtained is V/g. I substitute this into the trajectory equation (4), and we see that

$$(5) \qquad x_{max} = \frac{V^2}{2g}$$

We say, well, what more can we want? I can make the problem more complicated by assuming that we're going through an inhomogeneous atmosphere, so that in addition to the influence of gravity, we have a retarding force due to the velocity. If I make this a function of x and x' which is sufficiently complicated, I cannot solve it explicitly. But, we say we don't care. Give us any function $g(x,x')$, so that the differential equation is

$$(6) \qquad x'' = g(x,x').$$

I have the initial conditions. I can run out the trajectory, find out where the maximum time occurs, and get the solution to the problem.

What's wrong with that? First, perhaps it isn't fair to ask what's wrong with it; but let's ask, what it is we don't like about it. One thing we don't like is that this solution gives us too much information. Remember, I just asked for one bit of information, i.e., what is the maximum altitude? I'm not interested in the whole trajectory, simply the maximum trajectory. Furthermore, I would like to know what is the maximum trajectory as a function of the initial velocity. In order, therefore, to solve that problem, for every initial value V, you would have to run out a trajectory, assuming that you're in

the general case where you cannot solve explicitly, and from each one of these runs, extract one point, the maximum altitude. Then you would turn out with a curve.

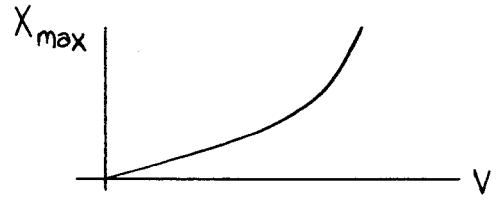


Fig. 7

But each point would require the calculation of one complete trajectory. This is a rather inefficient way of doing it.

Is there any way in which we can get an equation for x_{max} as a function of the velocity directly? Can we use another approach? We find there is another approach, and this is the approach that we are going to refine to the dynamic programming approach when we introduce maximization.

Let's start all over again and find the maximum height. We write down the following obvious statement:

<u>The maximum height depends on the initial velocity</u>.

Surely we all agree about that. It "depends on" is a mathematical translation of "is a function of", so I write

Maximum height = f(V).

I want to obtain an equation for this maximum height.
I observe the following:
I start the stone at altitude 0.

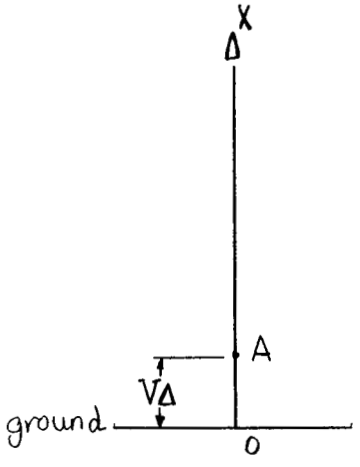


Fig. 8

At the end of the time Δ it has achieved a certain altitude, point A. Take Δ to be an infinitesimal, which means that the altitude it has obtained is $V\Delta$, if it's thrown up. Now, I see that whatever the maximum height starting from 0 was, it will be the altitude $V\Delta$, plus the maximum height obtained starting from A with a new initial velocity. Therefore:

> The maximum height starting at ground zero with velocity V is equal to the height attained in time Δ, plus the maximum subsequent height.

Let's translate this into algebra. We've assumed that the atmosphere is homogeneous, so we now have a new problem which we start from point A with a new velocity. What is the new velocity? We've lost $g\Delta$ due to the pull of gravitation; according to our definition of the function f,

f is the maximum height when we start with the velocity $V-g\Delta$ and this is all $0(\Delta^2)$, where is an infinitesimal. Our equation is

$$(8) \qquad f(V) = V\Delta + f(V-g\Delta) + 0(\Delta^2).$$

Equation (8) says that if I look at the process, after a certain time Δ has elapsed, I have exactly the same type of process, except I've started with a new velocity because I'm assuming, in this simple model, a homogeneous atmosphere. Let's expand in powers of Δ and let Δ approach zero. We're left with

$$(9) \qquad f'(V) = \frac{V}{g}, \quad f(0) = 0.$$

The initial condition arises because if the velocity is zero the maximum altitude is zero. The solution is

$$(10) \qquad f(V) = \frac{V^2}{2g} \qquad \text{the desired result.}$$

The nice thing about this approach is that we've found out only the information that was desired. We don't answer such problems as where the stone was at the end of two or three seconds. The question was: what is the maximum height given the initial velocity? This is what we answer. I make an issue about this because I want to emphasize that if you understand this problem, you understand everything that follows, for this has everything in it. It has the whole idea that we're going to use. The only thing that's going to be more complicated is that we're not going to allow the stone to follow it's own desires, or the pull of gravity, but we're going to determine what the velocity is going to be, as we go along.

So we're going to add some minimization and maximization. But the basic idea of looking at the problem in this way, translating an obvious verbal statement into an equation, is all that we're going to use, no more, no less sophistication.

Let us now consider a slightly more difficult problem:

> A stone is thrown straight up with velocity V into an inhomogeneous atmosphere with air resistance dependent on altitude and velocity. What is the maximum altitude it attains?

We are at ground zero, but as we go up to various altitudes, we find that there are strata. The strata have different densities, for example, so that we have a resistance which depends upon the altitude and, of course, upon the velocity. Now the classical approach is straightforward. I have the equation

$$\ddot{x} = g(x,\dot{x}), \text{ with } x(0) = 0 \text{ and } x'(0) = V \; . \tag{11}$$

As g is a function which does not yield to an explicit solution, all I do is run out the trajectories on the computer, and to each value of V, I get an x_{max}. This is the method that is used at the present time. We can do something which is a little better. Can we find the function of V directly? We can't do it easily in this case because of the inhomogeneity; so we have to extend the problem a little bit. Let's take the following more general problem:

Suppose I start an altitude h, and I throw the stone straight up. Then what is the maximum distance attained above ground?

This is the problem which we'll study. The maximum distance above ground depends clearly upon the initial distance h, and upon the velocity V. Let

$H(h,V) = f(h,V)$, + h = maximum distance above ground starting at altitude h with Velocity V, where f(h,V) is the maximum additional distance gained, which also depends upon h and V.

Now, what equation do we get? Just as before, I say let the process operate in infinitesimal time Δ. I started at h, I go up another distance $V\Delta$. Now, I'm in exactly the same situation, except my new distance above ground is $h + V\Delta$, and my new velocity is $V - g(h,V)$. My assumption was that in x", the acceleration was a certain function of position and velocity. This is $\Delta \cdot g(h,V)$. The equation is

$$(12) \qquad H(h,V) = h+V\cdot\Delta + \left[f(h+V\cdot\Delta,\ V-\Delta\cdot g(h,V)\ )\right] + O(\Delta^2).$$

Now if we let Δ approach zero, we get to a partial differential equation,

$$(13) \qquad 0 = V+V\frac{\partial f}{\partial h} + g(h,V)\frac{\partial f}{\partial V}\ ,\ f(h,0) = 0$$

where the initial condition arises as before.

Now if we wanted to pursue this analytically, we could

use the theory of characteristics. If we wanted to pursue it computationally, then we could solve this as a first order partial differential equation. Or, we can use the recurrence relation (12) itself.

This problem also requires a computational solution, and you might wonder what you have gained over the original computational solution. The answer is that if you solve this computationally, every number that you grind out is meaningful, because every one of these functions is an interesting function to the engineer or the aerodynamicist who is doing the problem. He wants to know how much farther you go if you're at a certain altitude and a certain velocity. Whereas, if one uses the conventional approach, he has to compute all the trajectories and pick out just one of them. So this function gives us the information we want in terms of the variables which describe the solution.

Let us now apply this technique to the calculus of variations. Before applying it to a general problem, let me apply it to a very simple problem, a problem of geodesics. Suppose, abstractly, I have a point p in phase space, and I want to go to another point r in phase space. In other words, for a three-dimensional trajectory problem

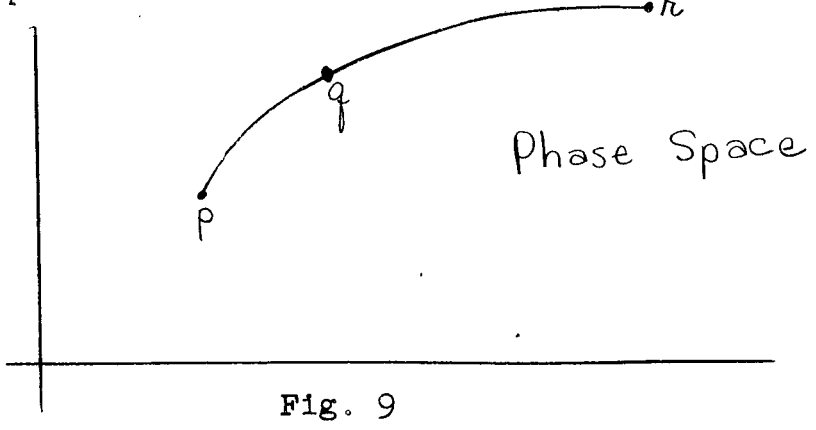


Fig. 9

this would be as follows: starting at a certain point with certain initial velocities, what's the minimum time required to get to another point with other velocities? For this, p may be two-dimensional, four-dimensional, six-dimensional, or may have higher dimensionality, depending upon the problem. When we say, p is a point in phase space, we mean p is essentially a finite dimensional vector whose components describe the state of the system. If it's a conventional problem in mechanics, then p has as components the position and velocity values. We want a path of minimum time. What can we say about a path of minimum time? Using the same idea that we did before, we say, suppose we continue along the path to some point q. We don't know at the moment what q we went to, but we can say that the remainder of the path must also be a geodesic. If pqr was a geodesic, say a path of minimal time, then, if we take intermediate point q, the path qr must be a path of minimum time. This is just the idea we were using before. We were saying that if we are throwing a stone straight up and are looking for the maximum distance above the ground, after we've gone a certain way along, we find exactly the same type of problem ahead of us. That's what we're saying here. The original problem was to find a minimum time from p to r, then at point q the problem must still be to find a minimum time from q to r. Subsequently, we'll describe why it isn't always true that the time from p to q is minimal. You might think that. It's true for some geodesics, but not for others.

As before, introduce a function.

$f(p)$ = minimum time to go from p to r, which depends upon where we start.

It's also a function of the terminal point, but for this discussion let's keep the terminal point fixed. Then $f(p)$ will be the time required to go from p to q, whatever that time is in our co-ordinate system, plus the minimum time required to go from q to r. This is a function which we've defined for all points p.

How should q be chosen? I say q should be chosen to minimize

$$(14) \qquad f(p) = \min_{q} \left[t_{(p,q_1)} + f_{(q_1)} \right].$$

Now there's one further idea added. If we have a choice of several q's, say q_1, q_2, $q_3 \cdots$,

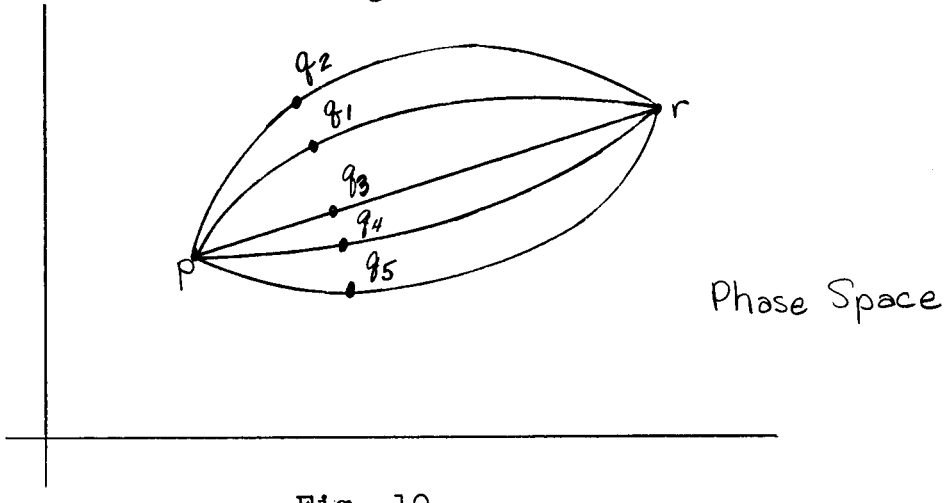


Fig. 10

how do we choose the appropriate q? We have to balance two factors. We have to balance the time to get to q_1

plus the minimum time to go from q_1 to r. So this part is analogous to what we did before. This is now a new idea. But it's again an obvious common sense point of view.

Equation (14), as it turns out, contains most of one-dimensional classical calculus of variations. I'll give you some references to this subsequently. This common sense, rather simple approach will contain the classical Euler condition and other conditions. However, we're interested in an approach which leads to a feasible computational method.

Let me give you a very simple illustration of this, which is quite pertinent to general trajectory problems. The problem actually arose in the following way: A friend of mine was traveling across the country in a plane, and he got into a storm. The pilot deviated from course (he announced that he was deviating); and afterwards my friend asked the pilot, "What rule do you choose to deviate?" The pilot replied, "Well, I have to fly within a certain distance of certain air fields." Since my friend was a mathematician, he immediately conjured up the following problem: He said, "Suppose we have a set of points on the map---cities, air fields--- which are numbered in some way

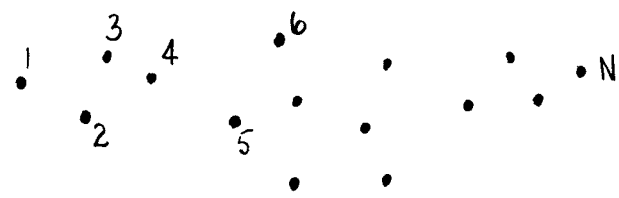


Fig. 11

with terminal point N. I give you a matrix $T=(t_{ij})$, where t_{ij} is the time required to get from the i-th point to the j-th point. I assume that any two points are connected, which, of course, in general, will not be true. All that means is that $t_{ij} = \infty$ if there's no trajectory between i and j. Computationally, you set it equal to a very large number, which effectively rules out the use of it as a path. I want to go from the initial point 1 to the terminal point N.

I can go in the following ways: I can go directly, or I can stop once and then go, or I can stop three times, four times, etc. This is a practical problem as far as the routing of traffice is concerned, because, if this is time, time is not directly proportional to distance, as we know in going through traffic; and very often we're willing to go several blocks or several miles out of our way, so as to minimize the time required to get from one point to another. So if you were going from one point to another in a city which you assume was laid out in a rectangular grid, an appropriate problem would be: Which streets do you follow at various times of the day in order to minimize the time to get from one point to the other?

This looks like a very combinatorial problem. It doesn't look like the kind of problem that one can handle by calculus or the calculus of variations, but it lends itself very nicely to the foregoing approach. The first thing I do is consider the general problem, not of getting from the fixed point 1 to the fixed point N, but of getting

from an arbitrary point i to the terminal point N. I can define a function

f_i = the minimum time required to go from i to N, i = 1, 2, . . . , N ;

f_i is exactly the $f_{(p)}$ that I defined before. I drew the geodesic, Fig. 10., as if it were a nice continuous curve, but, of course, I didn't define what my phase space was. This is a particular realization of the problem I was talking about before. And I'll explain shortly the relevance to actual problems in trajectory optimization. If we're at i, what are we going to do? Clearly, we must go to some other point j, so we use up the time t_{ij}. Then, starting from j, we go to N, and now we want to minimize

$$(15) \qquad f_i = \min_{j \neq i} \left[t_{ij} + f_j \right]$$

where we put $j \neq i$ because we insist that we go someplace. Also

$$(16) \qquad f_N = 0$$

which is a condition which gives us the unique solution. It's clear you can add a constant to both sides of the equation as it stands; but, if we add this condition, the time required to get from N to N, or zero, then it's not difficult to show we have a unique solution.

Computationally, how would we go about getting it? We have the unknown value on both sides of the equation, so we approximate in several ways. One way to approximate is what one might call policy space.

What policies could you have? The simplest policy would be always go directly to N. If you always go directly, $f_i^{o} = t_{iN}$, the time required to go from i to N. But what's a slight improvement on that policy? It would be to stop at one place in between. If you stopped at one place in between, and then went directly, what you would want to do is minimize over the point at which you start, and hence,

$$(17) \qquad f_i^{(1)} = \min_{j \neq i} \left[t_{ij} + f_i^{(0)} \right] .$$

If you continue in this way, it's first of all clear that your sequence is decreasing, and furthermore, you can show that it terminates at the end of a finite number of steps. So we have a very simple way of solving this problem. Subsequently, when I discuss the use of the digital computer, I'll discuss the feasibility of it. It follows that if you have, say, a hundred points, this is the type of calculation that one could do by hand in the space of a few hours. We tried it out on people who have had no mathematics at all, merely asking them to perform these very simple operations---adding, taking a minimum---it takes just a few hours to solve very big networks.

The relevance to actual trajectory problems can be shown by means of a one-dimensional problem. I start at $x(0) = c$

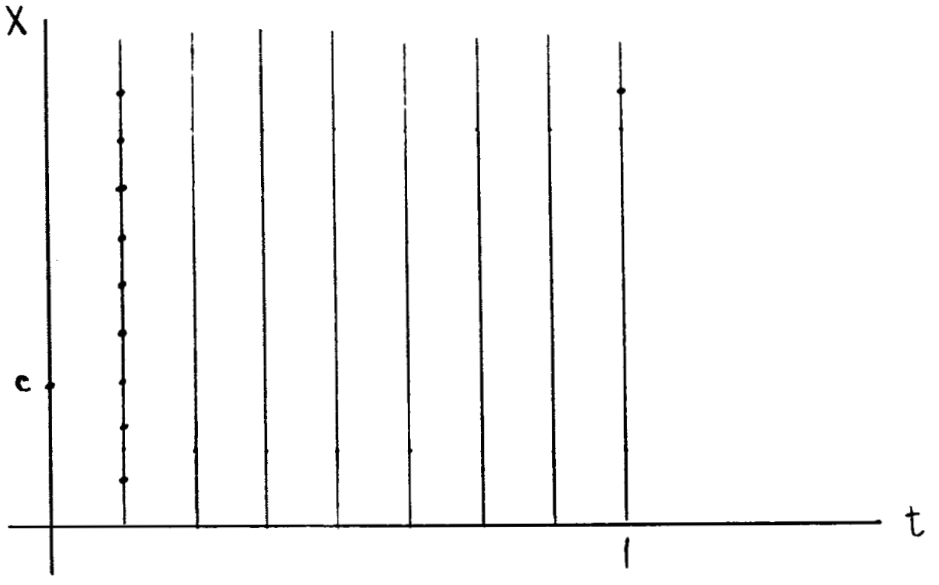


Fig. 12

and I want to end at $x(1) = X$. What we can do is approximate the problem in the following way. Let's just draw a vertical series of lines in $[0,1]$ and assume that in some way or another you can only be at a certain set of points on these lines. Assume that these points were all close enough together, so I know how to get to nearby ones. Instead of allowing an arbitrary step, assume that once we are on a given point on a given line we can only go to one of the nearby points in phase space. Then, by discretizing the problem in this way, we're back to this type of problem considered earlier, which we know how to solve. So the fact that this problem can be solved for very large numbers of points, as I will discuss later---several thousand, five, ten thousand points---means that we have a very quick way of approximating to the solution of quite complicated trajectory problems. I haven't established the feasibility, because I haven't discussed exactly how large N can be or what the time required is.

There's one other point I wanted to mention. I have said that I would show why the initial part pq of the optimal curve in phase space was not a geodesic. I said, suppose we want to go from point p to point r in phase space. It's clear that segment qr must be a geodesic. It's not at all clear that segment pq must be a geodesic, and, in general, it's not true.

When is it obviously true? It's obviously true for the following simple case: Suppose I take a certain point a,b in the xy plane,

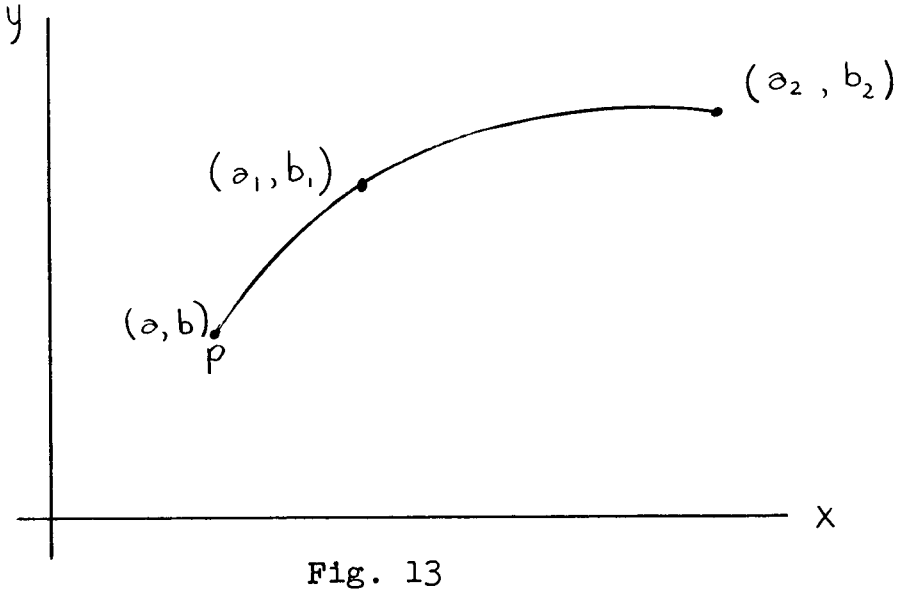


Fig. 13

and I want to go to another point (a_2, b_2) in minimum time. It's clear that both parts of the curve a, a_1, a_2 are geodesics, if the curve itself is a geodesic. The proof is by contradiction. If part a, a_1 were not a geodesic, I would use a minimum time path over a_1, a_2.

Suppose that, not only do I want to get to point (a_2,b_2) in minimum time, but I want to arrive there at a certain angle. In other words, I want a path which comes in at a certain angle. This is now my definition of a geodesic. Phase space is now not only in the position coordinates but also velocity or angle coordinates. Then it's clear that any terminal part of the curve must be a curve which comes into this point at this angle. But any former part of the curve is not a geodesic in this sense. So the recurrence relation, the functional equation that we use, works in the x-increasing direction if we take the end. But it holds only for subsequent times. We cannot make any statement in general about this.

This is quite different from what holds in the classical calculus of variations. In the classical calculus of variations when we have an optimal trajectory what we know is that any part of this trajectory is a solution of the Euler Equation. This information, which is quite useful, is also quite dangerous, because we know the Euler Equation can have many solutions. If it had a unique solution this would be very useful. If it had a multiplicity of solutions it doesn't give us the whole we want. When we remove the multiplicity of solutions, as we have here, we are saying from here on, that the solution will be not only a necessary condition, but will be characterized uniquely by the condition that it is an absolute minimum. We will discuss these things in more detail. I'd like to point out that we automatically get rid of the problem of determining what the absolute minimum is once we've determined various relative minima.

References

For the approach to the ideas of dynamic programming given here:

R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton U. Press, 1961.

For the connection between dynamic programming and the calculus of variations:

S. Dreyfus, "Dynamic Programming and the Calculus of Variations", Jnl. Math. Anal., Appl., 1960.

Calculus of Variations - Computational Aspects

by

Dr. Richard E. Bellman

Calculus of Variations -- Computational Aspects

In considering the minimization of a functional by means of the classical theory of calculus of variations

$$(1) \qquad \min_{u(a)=c} \int_a^b g(u, u', t)\, dt,$$

we face the problem of solving a problem given in terms of the Euler equation and associated end or natural boundary condition. The minimizing arc, of course, is from a class of arcs which satisfy the initial condition

$$(2) \qquad u(a) = c.$$

The initial point is fixed. The difficulties of this approach have been pointed out earlier. It is clear that the minimum value of the functional depends on the initial point a, as well as the initial value of the minimizing function c. We translate this mathematically, as before, by introducing the function

$$(3) \qquad f(a, c) = \min_{u(a)=c} \int_a^b g(u, u', t)\, dt.$$

Let the curve, Fig. 1, represent the minimizing arc.

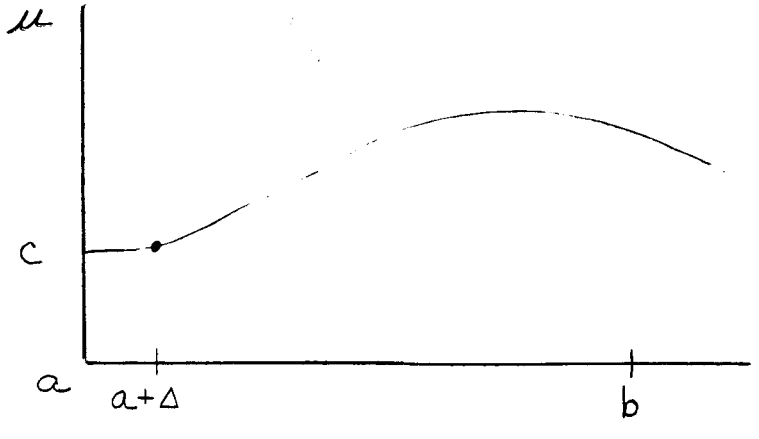


Fig. 1

If we apply the method developed earlier to this problem, we see that, for an arbitrary point $a+\Delta$,

$$(4) \qquad \min_{u[a,\ b]} = \min_{u[a,\ a+\Delta]} \left[\min_{u(a+\Delta,\ b)} \right].$$

In other words, no matter how we get to $a+\Delta$, the remaining portion of the curve over the interval $[a+\Delta,\ b]$ must be a minimizing arc.

It is important to point out at this point that there is a duality between the dynamic programming approach introduced here and the calculus of variations approach discussed earlier. The calculus of variations considers the minimizing arc to be a locus of points, and attempts to find it by solving differential equations. The theory of dynamic programming regards the extremal as an envelope of tangents, and attempts to determine the optimal direction at each point on the extremal. The former theory cannot be extended to feedback control and other problems; the latter can be extended naturally to include stochastic and adaptive control elements, since at each point, at each stage of the process, it gives as instructions the optimal direction in terms of present position.

Let us rewrite (3) using (4).

$$(5) \qquad f(a,\ c) = \min_{u[a,\ b]} \int_a^b g(u,\ u',\ t)\ dt =$$

$$\min_{u'(a)} \min_{u(a+\Delta,\ b)} \left[\int_a^{a+\Delta} g\,dt + \int_{a+\Delta}^{b} g\,dt \right]$$

Here we have already put $\underset{u(a,\ a+\Delta)}{\text{Min}} \longrightarrow \underset{u'(a)}{\text{Min}}$ as $\Delta \longrightarrow 0$.

Expanding in powers of Δ, where Δ is infinitesimal, if $a < b$,

$$(6) \qquad f(a, c) = \underset{u'(a)}{\text{Min}} \left[\Delta \cdot g(c, u'(a), a) + f(a+\Delta, c+u'(a) \cdot \Delta) \right] + 0(\Delta^2).$$

Finally, expanding once more and letting $\Delta \longrightarrow 0$, we obtain

$$(7) \qquad 0 = \underset{u'(a)}{\text{Min}} \left[g(c, u'(a), a) + \frac{\partial f}{\partial a} + u'(a) \frac{\partial f}{\partial c} \right].$$

For further details and particular references, see Applied Dynamic Programming by R. Bellman and S. Dreyfus, Princeton University Press, Princeton, N.J., 1962.

To illustrate the process, consider the problem

$$(8) \qquad f(a, c) = \underset{u(a)=c}{\text{Min}} \int_a^b (u'^2 + u^2 + u^4)\, dt.$$

This can be transformed by means of the above algorithm to

$$(9) \qquad 0 = \underset{u'(a)}{\text{Min}} \left[u'^2(a) + c^2 + c^4 + \frac{\partial f}{\partial a} + u'(a) \frac{\partial f}{\partial c} \right].$$

By ordinary differential calculus, we minimize over the function $u'(a)$, and we have the problem

$$(10) \qquad \frac{\partial f}{\partial a} = \left(\frac{\partial f}{\partial c} \right)^2 \Big/ 4 - c^2 - c^4, \quad f(b, c) = 0,$$

where the boundary condition follows from the definition

of f. This problem can be solved either by standard numerical techniques or by equally standard methods for partial differential equations.

Problem (1) is a very nice, but artificial, problem because there are no constraints on the minimizing arc. To make it more realistic, we add constraint of the form

$$(11) \quad |u'| \leqslant K$$

Equation (6) becomes

$$(12) \quad f(a, c) = \underset{|u'(a)| \leqslant K}{\text{Min}} \Big[g(c, u'(a), a) + f(a+\Delta, c, u'(a) \cdot \Delta) \Big],$$

where, by assuring that Δ is small, we can drop terms $0(\Delta^2)$ and smaller. Analytically, the addition of a constraint makes the problem more difficult; computationally, it is easier to solve. To illustrate the last remark, consider how one might solve the problem computationally. We first divide the time scale into sub-intervals of convenient size. Of course, the accuracy of the result will depend on the mesh size; but for the sake of argument, let it be of length Δ, a small but finite number. We can think either of taking steps $a \longrightarrow a+\Delta \longrightarrow a+2\Delta$, ... or b, $b-\Delta$, $b-2\Delta$, ... Since we know $f(b, c) = 0$, let us start at $t = b$.

Knowing $f(b, c) = 0$, take $a = b-\Delta$ and replacing u' by v

$$(13) \quad f(b-\Delta, c) = \underset{|v| \leqslant K}{\text{Min}} \left[g(c, v, b-\Delta) + f(b, c + v\Delta) \right]$$

The last term in the brackets must vanish in view of the end condition $(10)_2$, i.e. $f(b, c) = 0$ for any c. Computationally, (13) can be solved by a simple enumeration process which produces, for a fixed value of b, a table of $f(b-\Delta, c)$ for a range of c and v.

The role of the constraint, of course, is to allow us to scan a much smaller range of v than would be necessary without imposing a constraint.

The next step is to choose $a = b - 2\Delta$. The equation (12) becomes

$$(14) \qquad f(b - 2\Delta, c) = \min_{|v| \leqslant K} \left[g(c, v, b - 2\Delta) + f(b - \Delta, c + v\Delta) \right]$$

where we know a set of values for the last term in the bracket from (13).

Graphically, Fig. 2 in an a - c plane, we are given $f(b, c)$ along the line $a = b$.

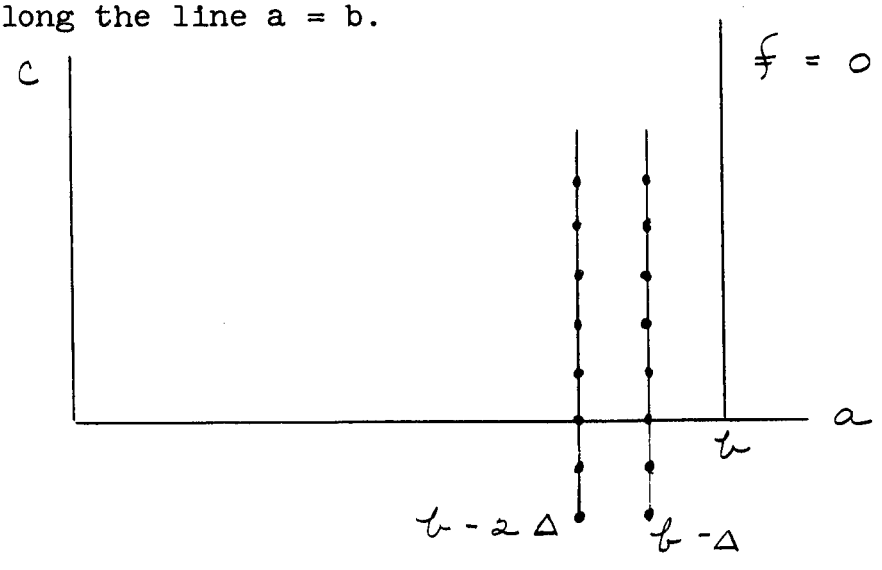


Fig. 2

The computation proceeds as follows: choose a grid of values for c, say, $c = k\Delta$, $k = 0, 1, 2, \ldots, R$. Using (13) we then compute a sequence of values of $f(b - \Delta, c)$ along the line $a = b - \Delta$. Step-by-step we build up a complete set of values of $f(a, c)$ to complete the calculation.

Here we can define the word policy $\equiv v(a, c)$. In a u - t plane

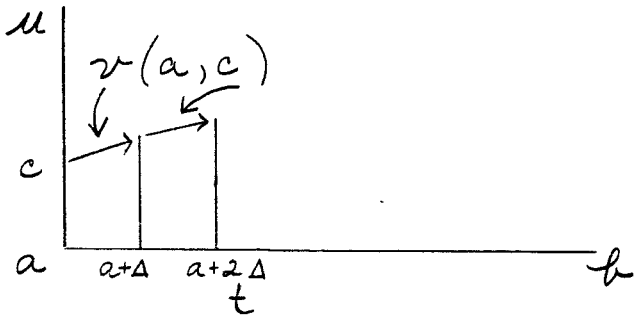


Fig. 3

the "policy" is the choice of $v(a, c)$, i.e. the slope of the minimal arc, at each stage of the process of marching out a solution to the problem of optimizing the integral (1).

By means of this simple algorithm a whole set of values of $f(a, c)$ is found; one can in principle obtain from this information a sensitivity analysis, information which is usually more important in engineering applications than just the one answer provided by the classical approach. One actually can't be sure that the one answer from classical analysis is the only answer or the answer sought. On the other hand, this simple computational method circumvents questions of continuity, differentiability, uniqueness, and, in particular, stability problems associated with finite difference methods for solving equations such as (10).

As an example of an instability which arises in even a very simple case consider

$$(15) \quad \underset{u(0)=1}{\text{Min}} \int_0^T (u'^2 + u^2)\, dt \quad , \quad T \text{ fixed}$$

The Euler equations are

$$(16) \quad u'' - u = 0 \, , \quad u(0) = 1 \, , \quad u'(T) = 0$$

The solutions are of the form e^t and e^{-t}. If one tried to solve this equation numerically regardless of the mesh size, Δt, round-off would introduce influences $O(e^t)$ and after just a few steps the solution would be dominated by the e^t term.

In the method of dynamic programming we are not interested in finding the solution, i.e. the locus of points forming the optimal trajectory over and over again for each initial value. We just want to solve the problem in the most direct and efficient way and get to the end point optimally. If we make a slight error along the way, it doesn't matter, for we are interested only in the part of the trajectory that remains. An error means we merely must make a slight change in direction on the next step. The process has feedback aspects and hence is inherently stable.

The above computational scheme is feasable on modern digital computers considering the size of current rapid-access storage. Slow access storage, e.g. magnetic tape, is ignored in view of the large retrieval

time. For example, an IBM 7090 can store approximately 32,000 ten-digit numbers. If our problem, equivalently stated, is

$$(17) \qquad f_n(c) = \operatorname*{Min}_{[v]} \left[g_n(c, v) + f_{n-1}\left(T(v, c) \right) \right]$$
$$u = 1, 2, \ldots\ldots, \quad T = v + c \cdot \Delta \text{ with } f_0(c) \text{ known}$$

then, for a reasonable grid size on the c - scale, say, $c = m\delta$ ($|m| \leqq M$), we must compute $2M + 1$ values of c. At each c, the values of $f_0(c)$ and $g_0(c)$ must be stored. For the minimization process we might scan the values of $|v| \leqq k$, for each value of which we must store $f_1(c)$, $v_1(c)$, etc. Thus at this stage we require storage capacity of the order of $3(2M + 1)$ locations. Of course, by various tricks one can store up to 100,000 ten-digit numbers, but let us disregard that in this estimate. A little arithmetic soon convinces one that the active storage of any modern computer can be exceeded in any moderate problem. This is the limitation on the computational feasibility of the method.

On the other hand, regardless of the complexity of the intregrand function in

$$(18) \qquad \operatorname*{Min}_{u(a)=c} \int_a^b g(u, u', t)\, dt$$

or the constraint functions,

$$(19) \qquad R_i(u', u, t) \leqq 0 \qquad i = 1, \ldots\ldots, K$$

the time required to make typical calculations by means of

the corresponding functional equation is the order of minutes even for a very fine grid.

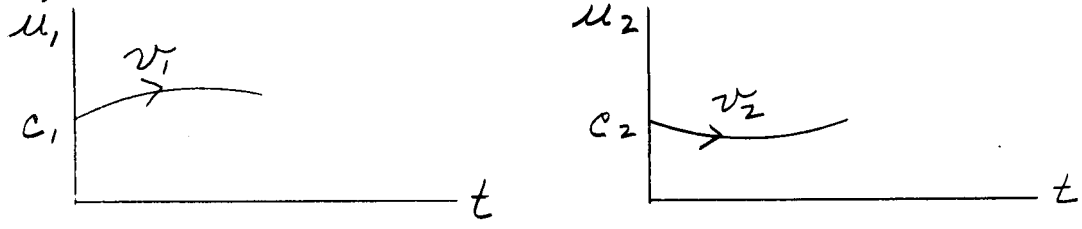
For the problem

$$(20) \qquad \underset{\substack{u_1(a)=c_1 \\ u_2(a)=c_2}}{\text{Min}} \int_a^b g(u_1, u_2, u_1', u_2', t)\, dt$$

the above results carry over directly if, now, all terms are interpreted as vectors, e.g.

$$(21) \qquad c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

The choice of a policy becomes the choice of two directions,



one in u_1 - t plane, and one in the u_2 - t plane. The computational feasability again depends on the rapid-access storage capacity of the digital computer. If a mess of N x N points is chosen in the c_1- c_2 plane, we have N^2 grid points. If, for example, N = 100, a not unreasonable number, $N^2 = 10^4$, which is already the order of the largest storage facilities currently in use. We see the fundamental difficulty is the "Curse of Dimensionality."

The amount of space required to store a function for values

of the argument can be reduced by certain techniques. Let us consider, in particular, the method of polynomial approximation of a function of one variable f(c). Every well-behaved function can be approximated in the form

$$(22) \qquad f(c) \cong \sum_{k=0}^{N} a_k \phi_k(c)$$

for example, as a polynomial approximation

$$(23) \qquad f(c) \cong \sum_{k=0}^{N} a_k c^k$$

or, as a solution to a differential equation

$$(24) \qquad f(c) \cong \sum_{k=0}^{N} a_k e^{\lambda_k c}$$

If we are told for example that polynomial approximation is used, we must store the vector $a = [a_o, a_1, a_2, \ldots, a_N]$. To recreate the function for a given value of c, then, we must, by the naive approach compute $c, c^2, c^3, \ldots, c^N$. This involves $2(N - 1)$ multiplications. Of course, the number of multiplications can be reduced by writing the polynomial as

$$(25) \qquad (\ldots (((a_N c + a_{N-1}) c + a_{N-2}) c + \ldots\ldots + a_o)$$

which involves only N multiplications.

In the more general case, assume that we will store the function in the form (22). To save storage, we must pay in time for recreation. Thus we seek $\phi_K(c)$ that

are easily computed. If $\phi_k(c) = c^k$, $-1 \leqq c \leqq 1$ least squares approximation

$$(26) \qquad \min_{a_k} \int_{-1}^{1} \left(f(c) - \sum_{k=0}^{N} a_k c^k \right)^2 dc$$

leads to system of linear algebraic equations of higher dimensionality than is convenient to solve. On the other hand, it is easier to make use of orthogonality properties of special polynomials - they are still polynomials - that have the advantage of satisfying simple three term recurrence. For example, one could replace c^k by $P_k(c)$, the Legendre polynomials. Then

$$(27) \qquad \min_{a_k} \int_{-1}^{1} \left[f(c) - \sum_{k=0}^{N} a_k P_k(c) \right]^2 dc$$

leads to the usual expression for Fourier coefficients

$$(28) \qquad a_k = \frac{2}{2k+1} \int_{-1}^{1} f(c) P_k(c) \, dc$$

By making use of such methods, the time required to solve a four dimensional problem of the type discussed above can be reduced to about two hours per stage.

Dynamic Programming

and

Stochastic Control Processes

by

Dr. Richard E. Bellman

Dynamic Programming and Stochastic Control Processes

After having discussed the calculus of variations and illustrated how dynamic programming techniques may be used to treat problems in this area, let us discuss the same subject in rather an abstract way. This will be useful because we want to turn to the study of stochastic control and adaptive control; if we see a topic in a more general, more abstract setting, it makes it easier to extend the techniques that we've been using so far.

Dynamic programming is really an advertising name for mathematical theory of multi-stage decision processes. It is rather interesting to ask, why the name dynamic programming? Why not something nice and respectable such as the mathematical theory of multi-stage decision processes? First of all, it's clear that for advertising purposes, the latter is too long to stretch across a page. But the real reason was that back in 1949 and 1950, when I was working on the problems which led to the development of this theory, I was thinking in terms of decision processes. I didn't like the term decision process itself because decision theory already existed and was tied in with the world of statistics. The problems of decision theory are particular cases of dynamic programming problems, but they're quite specialized. I thought about planning, but planning was definitely out for other reasons. Just about that time, programming had come in. Programming was a word which had no real meaning, so it was very useful. And, of course, if you wanted to be strict about it, programming really does mean thinking of a program, which is planning (making decisions). Finally I wanted an adjective to modify programming because there was another technique called linear programming with which I didn't want to get confused. It was necessary to emphasize the fact

that this was a multi-stage process that dealt with processes over time.

The classical word of mechanics for non-static, time-varying processes is dynamic. Dynamic programming sounded like something that everybody should do. So I used the name, and, as I say, it has the very nice advantage of seeming to mean something but meaning nothing; you can do anything you want under this guise.

Actually I got interested in the program and problems of this nature in connection with a quite specific problem. When I was a consultant at Rand in 1948, the Air Force was playing a very important role as a deterrent, and S.A.C. (Strategic Air Command) was the most important weapon that we had. The question was how to use S.A.C. in the most efficient way. In the beginning, people thought very naively in terms of one massive raid because, as usual, the generals and admirals were always fighting the last war extremely well, analyzing it, and writing their memoirs. As Churchill, I think, said about some famous general, "He sold his life very dearly to a publisher." This leads one to an analysis of the last war. When they thought about bombing raids, they thought about the bombing raids from London over Berlin, a distance of about four or five hundred miles, so that planes could come and go.

Eventually, by 1948 or 1949, somebody looked at the map and realized that with the ranges we then had it was impossible for planes to come and go, expecially in daylight, which meant that if you had a large number of targets, you had to think in terms of a multi-strike operation

in which you bombed a certain number of targets, a certain number of planes returned, you attacked some further targets, and so on. This had many complications. In the first place, it was a multi-stage affair. Secondly, it had stochastic elements in it. You couldn't really plan your second raid efficiently before you knew what happened on the first raid. Nonetheless, you had to decide what to do at the first raid. Some new mathematical features were definitely present.

Like many mathematical efforts in connection with new problems, I came out with some good techniques, although I never made any contribution to the original problem. Fortunately, time took care of that, as I mentioned earlier. This is the situation of most mathematical endeavors -- one does very interesting work, very nice theories come out of it, and fortunately, technology takes care of the actual problems. After I got interested in the whole field of multi-stage decision processes, I realized quite quickly that these problems were common to the field of engineering in the way of control theory, and in the way of calculus of variations. We had similar problems in the field of economics, operations research, medical diagnosis, and across the board. Multi-stage decision theory is one of the most important mathematical theories as far as applications are concerned. There are many mathematical problems, interesting for their own sake, in the field. So let me now discuss the thing in an abstract fashion, introduce a few bits of terminology, and then apply this technique to stochastic control processes.

We shall think abstractly in terms of the system.

The system may be a satellite, a decoy, a factory, or maybe a human being. A basic element of the theory is one system. The standard technique for studying a system mathematically is to introduce a way of describing it. Let us introduce a state variable, a vector p which usually depends on time t. For example, if we're talking about a satellite, one state variable would be its position and velocity at a certain time. It might be the amount of fuel, and/or several other factors. For simplicity, let's take time to be discrete -- t = 0, 1, 2, ... In this way we eliminate a number of spurious problems concerning continuity and differentiability, which allows the principal problems to come out rather clearly.

One of the difficulties with the classical calculus of variations is that an enormous amount of time is spent worrying about existence and uniqueness of various types of solutions, and very little time is devoted to such questions as whether you can get at these solutions if they exist. If we make time discrete, then we automatically bypass all problems of existence and uniqueness. We now are dealing with the minima or maxima of finite sets of quantities or with finite numbers of possibilities.

Clearly, one paramount difficulty is whether we have a feasible technique for obtaining solutions. This is <u>the</u> important mathematical problem if you're dealing with the physical world. Let us ask ourselves what happens to the system over time. Assume that this system is stationary, i.e., the mechanism doesn't change over time. The state of the system changes over time, but the transformation whereby the system goes from a state p to a neighboring

state p_1 is the same. Disregarding the decision and control aspects, let's look at the classical description of mathematical physics. If we define the system by means of a state variable, then we introduce a cause and effect relation. If the system is in state p at time 0, it will definitely be in state $p_1 = T(p)$ at time 1, where T is a given transformation. Now the study of the system over time is just the study of the iteration of this transformation:

$$(1) \quad p_2 = T(p_1) = T^2(p) \ldots.$$

This is classical analysis, and it stems from the ideas of Poincaré, Hadamard, Birkhoff, et. al.

We disregard the actual form of the equations for the moment and see if certain properties of the transformations lead us to certain conclusions about the behavior of the system. Classical mathematical physics, in many cases, becomes the study of the iteration of certain transformations. In this way, of course, you're led to the Ergodic theorem and fixed point theorems, etc. It's a very natural transition. Those of you who are interested in pursuing some of this may refer to the book I mentioned before on adaptive control processes, where you'll find some discussion of this, as well as some references.

Suppose we're interested in control theory. Not only are we interested in studying the evolution of the system over time, but because we're not satisfied with the evolution of the system over time, we will attempt to change it. How shall we portray that in an abstract fashion? We can think of control in the following way: The system is in

state p. Control means that we have a choice of the transformation that we can exert upon the system. But usually when we give something and take something, there is a certain cost. We may be able to get more control and, say, minimize the deviation from a desired state, but only at the expense of additional cost in resources or time. We have to balance the cost of control with the cost of deviation from desired state. Symbolically, given a set of transformations $T(p, q)$, state S: $p \longrightarrow p_1 = T(p, q_1)$, where q is the control variable.

For example, if we're talking about controlling a trajectory, at each particular point our transformation tells us where the particle will be at the end of a certain time. One can think of q as the direction that we choose at a given point. As a result of choosing q at one point, we end up at a certain other point at the end of a unit of time. So it's interesting to think of control theory as a choice of a transformation to exert on the system at each time.

It's also important to realize that the dicision to do nothing is a very important one. That's also one of the control variables. Many people, of course, think that they're doing nothing when they make no decision. This is a definite decision which very often is the worst thing that one can do. Of course, one has to account for the fact that very often you can be ruined by means of a theory. One must balance these two ideas. It's important to realize that, in many situations involving uncertainty, you can only be destroyed by a theory. If you did nothing and let yourself just be oscillated by random forces, you know that after a large number of steps you're only going

to be $O(\sqrt{n})$ from where you were initially, but if you have a theory, you could be $O(n)$ away.

We said control is the problem of choosing a transformation. We choose q_1 and the state variable p and we get $p_1 = T(p, q_1)$. At p_1 we choose another control variable and we get to $p_2 = T(p_1, q_2)$, and so on. The control process is then equivalent to a choice of the q_i at each stage.

The difference between a control process and an ordinary process in mathematical physics is the following: In the first place, we agree that we're just studying the behavior of the system, not trying to alter it. In the second place, we have no evaluation of the outcome of the system. We don't particularly care what happens one way or the other. In a control process, we have the criterion function

$$(2) \quad R(p_1, p_2, \ldots, q_1, q_2, \ldots).$$

We may want to get someplace as quickly as possible or subject to minimum fuel, or we may want to keep the deviation between what's happening and some desired state as small as possible, and so on.

We care what goes on in a control process. The criterion function (2) is some function of all the states and of the decision, which makes things much too complicated. To simplify the ideas, let's consider control processes of the following type: Assume that there's a finite number of stages and that the criterion has the following form:

$$(3) \quad R = g(p_N) + h(p_1, q_1) + h(p_2, q_2) + \ldots + h(p_{n-1}, q_{n-1})$$

where $g(p_N)$ is some function of the final state. This is sometimes called terminal control. In many situations you don't care what happens during the process. You only care what happens at the very end of the process. For example, with a certain amount of fuel, you may wish to get to Mars. You don't particularly care what path you take as long as you get there. Next, we assume that at each stage there's a cost for the state and one for control. A typical example of this would be an economic situation in which one is trying to meet a given demand. Suppose the demand curve is as shown in Fig. 1.

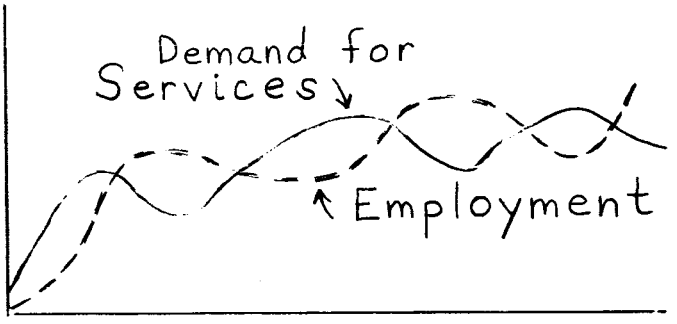


Fig. 1

Let your control variable be the number of men employed. You have a choice of hiring or firing men at each stage. You try to follow the demand curve with your services, but when you are above the demand curve, you must charge yourself with having people whose salaries you were paying, though they were producing nothing. When you are below, you have to charge yourself for, perhaps, buying supplies from a competitor at a premium price. So you have the situation where at each point you have two costs, one the cost of control, and one the cost of the deviation from

the state which you'd like to be in. This is typical of control processes. Abstractly, the problem is

$$(4) \quad \min_{[q_1, q_2, q_3, \ldots, q_N]} R$$

If once again we assume there are only a finite number of values of p and only a finite number of values of q, a finite number of states, and a finite number of control variables, there is no question of existence of the solution. We don't have to make any assumptions about continuity, for the problem is completely a finite problem. We have to take the smallest value of R from among, in most cases, a very large number of possible values. If the number of stages is large, and if the dimensions of q are fairly large, it's clear that we don't want to tackle this problem by enumeration. On the other hand, we don't want to tackle it by calculus because, in many cases, the functions are too complicated or the q's are discrete. For example, each q may just have the values +1 or -1. We must do something better. We have to have an algorithm which reduces a multi-dimensional problem to a sequence of lower dimensional problems. Looking upon the problem as a sequential problem, we want to make our decisions in sequential fashion corresponding to the actual control problem. The minimum depends upon two quantities, the initial state p, and the number of stages.

$$(5) \quad \min_{[q_1, q_2, q_3, \ldots, q_N]} R = f_N(p)$$

Of course, it also depends upon the forms of the functions, but they don't change. They are given to us. The only

things that change as we go along are the current state and the number of stages remaining. We could write the functional dependence as f(N, p), but we usually use discrete numbers as subscripts.

We shall now use the method of continuity comparatives. To illustrate the alternative methods, consider the following case: If you were studying religion or linguistics or anatomy, there are several ways of proceeding. First you can take an individual religion or individual language or an individual organism, and you can study this very completely, making a detailed, isolated study. On the other hand, you can take a family of related objects and trace the transitions as the organisms increase in complexity. It's very often much easier to understand a very complex organism as a limited form or as a sister form of other organisms than it is to understand the organism in isolation. This is the very important comparative method.

In mathematics, the method of continuity says that if you want to study a certain object, put it in a family of objects and go continuously from a member of the family that you understand quite well to the member which you don't understand initially as well. And by tracing the properties of the object in a continuous way, you can explain the properties of the desired object. This is, generally speaking, the imbedding technique. If you want to study a particular problem or a particular process, you must imbed it in a family of similar processes, and you do it in such a way that you have a very simple transition from a simple member of the family, which you understand, to the more complicated member.

In the present case, the simplest problem is the single-stage process. We bypass the multi-stage aspects, and find it is necessary to make only one decision. We wish to go stage by stage from the one-stage process to the two-stage process to the three-stage process, etc. Of course, the fundamental observation is that, in a process of this type, if we start out initially in an N-stage process, after one decision, we're going to be in an N-1-stage process. So we have a simple way of going inductively from N to N-1, a sort of backwards induction.

We must make some initial decision, the choice of q_1. That's going to cost us some function of the initial state and the initial decision. Now we have N-1 stages remaining, and we're in the new state $p_1 = T(p, q_1)$. It's clear that no matter what state we're in now, and no matter how many stages are left, we're going to proceed so as to minimize. This is just an extension of this geodesic property discussed earlier. The tail must always be optimal. This is worth dignifying under the name of the principle. I call it the Principle of Optimality. (Optimality is not a good English word, and therefore, one can use it freely.) The tail of an optimal policy must itself be an optimal policy with respect to the new state. It is the property we've used over and over again. Regardless of what q_1 is, what remains must be the minimum. We continue in a minimum fashion.

The question is how to choose q_1. We must choose q_1 so as to balance the cost incurred immediately, and the cost incurred over the remaining N-1 stages. Our functional equation is

$$(6)\quad \min_{q_1}\left[g(p_1) + h(p, q_1)\right] = f_1(p)$$

$$f_N(p) = h(p, q_1) + f_{N-1}(T(p, q_1)) \qquad N \geqslant 2 .$$

This is our general abstract formulation. We still haven't defined what the states are. They could be finite dimensional, infinite dimensional; they could be probability distributions, as they are in many cases; or we've said that, if a system is specified by a state, a decision is equivalent to a change of that state, and at each stage we incur a certain cost which depends upon the state and the decision which is made. We have decomposed a multi-dimensional problem into a sequence of one-dimensional problems. We make only one decision per stage.

Of course, this is telling us more than we ever wanted to know, because we only wanted to solve one problem. Now we have to solve a whole sequence of problems, not only a sequence of problems in N, but for arbitrary initial states. The answer to that is that most often in engineering problems you want a sensitivity analysis. The solution to (6) is just the information that is desired. It tells us how the minimum cost varies as a function of the initial state and the number of stages.

Let's introduce just two more terms. The <u>policy</u> is what we do in terms of where we are. The policy is a set of functions which tells us what we do in terms of the state and the time remaining. An <u>optimal policy</u> is a policy which optimizes. It provides the minimum cost or the maximum return, etc.

There is another interesting consequence of thinking

in terms of policies as functions. For example, suppose we want to compute the minimum time required to reach the origin from some point in phase space. In this case, we get an equation like (6). The minimum time is the time required to go someplace as a result of the first decision, plus the minimum time required to go from the new point. This is the geodesic property. If we have an equation like (6), we don't have the recurrence property that we had before. So far, we have spoken in terms of problems where we started with a known function. We then used the functional equation to get the second and third functions, and by simple repetition of iteration, we arrived at a desired solution. Now we have the unknown function on both sides, as in the optimal routing problem discussed earlier.

There are two procedures to handle this case. One is approximation in function space, which proceeds in the following way: We guess an initial function f_1. Then we compute f_2. Next we iterate over the functions. This is what is known as successive approximations, but of a very special type. It's successive approximations in function space. We can proceed in another way, giving equal emphasis both to the return function and to the policy function. One determines the other. If we have the equation

$$(7) \quad f(p) = \operatorname*{Min}_{q} \left[h(p, q) + f(T(p, q))\right],$$

then a choice of q, $q(p)$ tells us how to proceed. On the other hand, once we have f, then q is determined as the function which minimizes (7). So there's a duality between the two. This is a more general version of this Euclidian duality which I mentioned before, the locus of

points - envelope of tangents duality. We approximate either in function space or in policy space. Approximation in policy space goes the following way: First I guess an initial policy $q_0(p)$. Then I determine the return from the policy.

$$(8)\quad f_0(p) = h(p, q_0) + f_0(T(p, q_0))$$

$$= h(p, q_0) + h(T(p, q_0), q_0) + \ldots.$$

In other words, I get the return by just iterating, assuming I'm doing the same thing at each stage -- picking q_0 to be $q_0(p)$. Next, we find q_1 as the q which minimizes

$$(9)\quad h(p, q) + f_0(T(p, q)).$$

This says I'm going to approximate in policy space in the following way: First I will pick a policy which I'll call A. I just apply A over and over again. If I know at a given stage that I'm going to apply A over and over again, which gives me a certain return, what should my best first policy be? The answer may be to apply B first. Then I ask what the best second policy is, knowing that I'm going to first choose B, followed by A over and over. The answer may be B. But now I return to the question, if I apply policy BBAAAA..., is there a better first policy, say, C? In this way, the initial approximation slides away to infinity, and you end up with an optimal approximation.

This is quite different from the usual approximation. You're not just approximating in function space, but improving the policy at each stage. It can be shown that

this gives a monotone approximation. Thus you can improve upon anybody else's policy at very least. This is an important point. In choosing q_1 so as to minimize (9), if we choose the value of q before choosing q_0, we get back our old function. If we choose the minimum q, we have something definitely less. I would like to emphasize the importance of the policy. It isn't so important in deterministic processes when one can use conventional representations, but for more complex processes where one might not be able to describe the problem in conventional terms, then a policy is still sensible. This leads to something I want to mention briefly without going into. Once we have described the policy, i.e., we have specified the type of control policy we're going to use, then, even though the concept of a return function may not be meaningful because we may not know cause and effect well, we can carry out either in real life or with analog or digital computers the simulation processes. We can ask what would happen if we used this policy or another one. Of course, this is just exactly what is done in real life situations all the time. People test out policies which are sensible even in situations in which they do not have any analytic formulation of the problem.

The concept of a policy trmendously extends the scope of mathematics. Conventional mathematical techniques will not carry over to most real life situations, but we can still construct simulation processes, think in terms of policies, and extend the scientific method, if not the purely mathematical method. A policy will very often be a simple thing, whereas the analytic solution may be complicated. This is more often the case than not. For example, in courses on differential equations, you have

the following problem: A rabbit is at R, and there's a dog at D. The rabbit is going in the x-direction with a certain velocity, and the dog is constrained, i.e., its policy is always to point at the rabbit. The question: What is the dog's trajectory? It is the curve of pursuit.

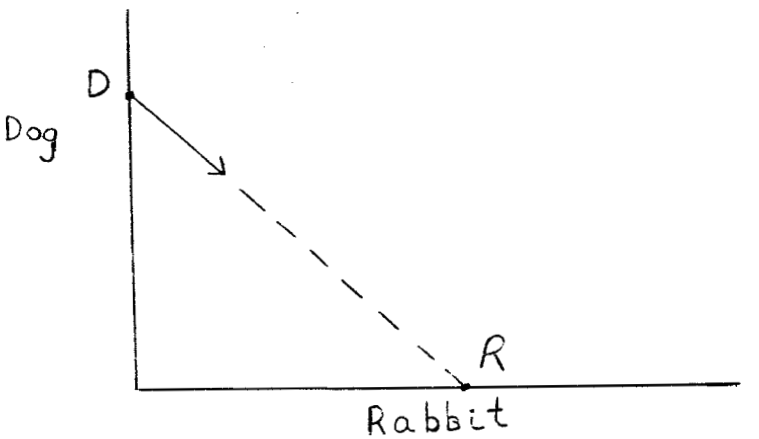


Fig. 2

This is an extremely complicated expression analytically. If you add a few more assumptions, you can get to the point where the differential equation cannot be solved explicitly. The actual analytic form of the curve can be quite complicated, but the policy is very simple. This simple pointing technique is used for actual interception problems.

Stochastic Control Processes

The word stochastic replaces the old word random. Random is a very bad word because the mathematical definition of random is almost exactly opposite from the definition implied by the ordinary English usage. To illustrate: When you tell a person you're doing things at random, it means there's no law, rhyme, or reason behind it. When a

mathematician says he's picking a number from the interval $[0, 1]$, he means he has a distribution, which is to say that if you pick a large enough sample you're going to see a great deal of regularity. So the English usage is quite different from the mathematical usage, and for that reason, for example, if you ask a person if 0.121212121 is a random number, he will say it isn't since it has too much regularity. Mathematically, of course, it's a perfectly good random number and has as good a probability of being chosen as any other nine-digit number that you can think of. Thus, instead of random, people use the word stochastic. Stochastic doens't occur very frequently in cocktail conversation. You can't turn to the girl next to you and ask, "Did anything stochastic happen to you last night?" She might slap your face!

Stochastic comes from the Greek word ΣΤΟΧΟΣ, meaning "a target". Shooting arrows at a target was a haphazard affair, and so from that you have the word stochastic. We deal with stochastic processes because we don't know how to make them deterministic. This is a point which isn't emphasized sufficiently. Probability is a very beautiful device for getting around ignorance. Naturally you would like to get rid of probabilistic considerations whenever you can. The simplest example of that is tossing a coin. Theoretically, when you toss a coin, if you know the angle of inclination and the force that you give it, the elastic properties of the floor, etc., you should be able to predict whether the coin will come up heads or tails. But it's a highly unstable situation. We know that the slightest difference in some of the initial conditions or the environmental properties will change a coin from falling heads to falling tails. Therefore, in situations like that, we

fall back on random variables. We assume that we have a coin -- heads is 0, tails is 1 -- and we have a certain probability p.

The meaning of this is completely clouded. Nobody has the faintest idea how to do probability in some satisfying way, unless you do it in an axiomatic way. When you try to make it sensible, you get into the following difficulties: What is meant by a coin that has a probability p of falling heads? It means that you keep tossing this coin a large number of times, and you get various sequences. If you toss it 10^6 times, there should be approximately p x 10^6 heads. But if you toss a coin 10^6 times, how do you know you're tossing it in the same way each time? After the coin has hit the floor 10^6 times, you don't have the same floor or the same coin. You're not performing the same experiment, which means that if you try to set up this concept of probability on a very common sense experimental basis, then you get into complete paradoxy. Suppose that even after you did it 10^6 times, you found 10^6-1 heads. Should we assume that the coin is very heavily loaded? Either it's very heavily loaded, or else I have a very unusual sequence from a fair coin. How do I know? That's a higher probability. You just keep pyramiding these difficulties. Consequently, people recognize all this. But let's start the whole thing over again from an axiomatic basis, exactly the same way that we do in geometry. We set it up on an axiomatic basis and leave it to the risk of the user as to whether he believes that any theorem which is sensible for a triangle that I draw on the blackboard is sensible for a triangle on the surface of the earth, etc.

The big problem in the use of probability is not

what you can do axiomatically and analytically, but whether the hypothetical situation has any correspondence with reality. Most of the time it doesn't! Most of the applications of statistics are completely spurious and open to many interpretations. You have to view them with the greatest possible care. If you are working for a cigarette company, you could easily get statistics that would prove conclusively that smoking is good for you. The Atomic Energy Commission, I'm sure, has experts who could produce testimony to show that fallout is good for you. You just pick and choose very carefully, and you can prove everything by means of statistics. You remember the famous remark of Disraeli, who said that there are three kinds of lies -- lies, damn lies, and statistics.

Consequently, I would like to warn you in advance that whenever you use probability theory, you're treading on very dangerous ground. You have to be very careful that the whole process is meaningful. You find people talking very blindly about the probability of war. This is a complete misuse of the concept of probability, either in practical terms or in the axiomatic sense. Probability is just thrown around very loosely.

We will introduce stochastic control processes from a purely mathematical, axiomatic point of view. Whether they have any relevance to anything that goes on in the real world is another matter. You are now forewarned as to the weaknesses and flabbiness of possible probability theory. I hate to disillusion you, but it's necessary that you know how dangerous it is to apply mathematical techniques to engineering problems. This is why experience and intuition, when combined with intelligence, are very

useful. Unfortunately, most of the time we have a dichotomy -- intelligence without experience, and experience without intelligence. Naturally, there's very little communication between the two. But as far as the engineering world is concerned, and certainly the economic world and more difficult worlds outside of those, mathematical methods should be used with the greatest care and caution, and nobody should ever take too seriously the results of analytical calculations.

One has to be very careful before he extrapolates from the many assumptions that go into writing equations to any realistic and complicated system, regardless of how many digital computers are there to testify as to the fact that these numbers were actually computed. The use of the digital computer is like the gun on the wall of the big game hunter. He says, "If you don't believe my story about shooting the rogue elephant, I'll show you the gun that did it."

Feedback Control Theory,

Stochastic Control Theory

and

Adaptive Processes

by

Dr. Richard E. Bellman

Feedback Control Theory

Today I want to talk about feedback control, starting with deterministic control processes and going on to stochastic control. Yesterday I warned you about the dangers, fallacies, and weaknesses of the calculus of variations. Although we would like to make use of it in computational schemes, usually we can only use it as a purely mathematical tool.

The concept of feedback control is a very interesting one and a very fundamental one. It is probably the most fundamental single scientific concept, because it cuts across the fields not only of engineering and economics, but also the fields of biology, medicine and psychology. More and more we're going to find it as one of the unifying concepts of science, and when we get more into mathematical biology and look into the functioning of living organisms, we're going to find that the feedback control concept, allied with this word homeostasis, namely the desire of the organism to keep itself the way it was, at the status quo. This is one of the guiding scientific principles. It is very interesting that mathematical techniques designed to handle quite specific problems in one field prove to have much wider validity and can handle problems in other fields which are superficially quite different, but abstractly and intrinsically exactly the same mathematical problem. For those of you who are interested in biology, psychology and medicine, it should be pointed out that the ideas that will be discussed in what follows have immediate application to these areas. And of course, as I mentioned at the beginning of the first lecture, in my opinion they're infinitely more important and interesting but, fortunately, science is a matter of taste.

Consider a system, the state of which we specify by a vector p at time t. Under time, p goes into a new state T(p) if it's uncontrolled.

$$\text{State S:} \quad p \longrightarrow p_1 = T(p)$$

If we exert control, then a system at state p under the influence of the control variable q goes into a state T(p,q). Abstract versions of this have recently come into some prominence. These are called sequential machines. Logicians are fond of them. There will be a certain flurry of interest in sequential machines - probably about 273 papers written - and then the field will die down, because there are no numbers attached to it. It would be interesting to make a count and see how accurate the number 273 is. Just in the last year or two people in logic have begun to realize that these general systems can be put into abstract format. But as I say, the weakness of that abstract format is they have not considered the numerical problems.

Consider a very simple one-dimensional case. Take a system S at time n, where time is discrete, and let the state be u_n $(n = 0,1,2...)$. Assume that at a time n + 1, the state is a certain function of the previous state and the control.

(1) $$u_{n+1} = g(u_n, v_n) \quad \text{where } v_n \text{ is the control variable.}$$

Suppose the system is originally in the state c. Everything is now one-dimensional. We can let all the variables be discrete if we wish, and as I say we can then avoid many of the sophisticated concepts of continuity, the question of existence, and of minimum and maximun values, and so forth.

As usual we assume that the criterion function is some function of the terminal state plus a sum of the costs incurred at each stage.

(2) $$R = K(u_n) + \sum_{k=0}^{N-1} h(u_k, v_k)$$

The cost incurred at each stage is a combination of some function of

the state variable and the control vector. Let's give an example of this. Suppose I have a linear system

$$u_{n+1} = au_n + v_n, \quad u_0 = c \tag{3}$$

where we start the system at state c. If our objective is to keep it in a state b, we can say that at each stage we have a cost of deviation $(u_k - b)^2$ and let's say we have a cost of control λv_k^2. We'll assume that the costs are additive, so our criterion function becomes

$$R = \sum_{k=0}^{N-1} \left[(u_k - b)^2 + \lambda v_k^2\right]. \tag{4}$$

We have no terminal cost.

This is a typical quadratic control process. It is an interesting one because it can be solved explicitly in several different ways; we may have time to come back to it later. Let's talk about the general case of which this is a particular example. Suppose our problem is to minimize the criterion function R over the choice of control v. I don't have to specify whether I'm using feedback control or whether I'm operating sequentially or whether I want to change all the v's at one time; it doesn't make any difference. Here we have a deterministic process. I emphasize this point now, because it will make a big difference when we talk about stochastic and adaptive processes. Here we can say we are concerned with a deterministic process; we can think of it as one n-dimensional problem where we choose v_0, v_1, . . . up to v_{n-1} at one time or we can think of it as a sequential process where we choose first v_0, then v_1, then v_2, and so on. If we use the dynamic programming approach, we say the criterion function

$$\min_{[v]} R = f_N(c) \tag{5}$$

is a function of the number of stages N and the initial state c and we get in the usual way

$$f_N(c) = \underset{v}{\text{Min}} \left[h(c, v) + f_{N-1}(g(c, v)) \right] \qquad N \geqslant 1 \tag{6}$$

$$f_0(c) = \underset{v_0}{\text{Min}} \left[h(c, v_0) + k\ (g(c, v_0)) \right].$$

This is a feedback control problem. Equation (6) is the dynamic programming formulation of it. We have reduced the problem to a sequence of one-dimensional problems rather than one n-dimensional minimization problem. The advantages of this are that we very often have a superior analytic technique; in most cases we have a superior computational approach. We don't have to let v be continuous; we can take $v_0 = \pm 1$. That actually cuts down on the computational time. We can thus take advantage of all kinds of realistic constraints, realistic functions, and we do not have to tailor them so that they obey some particular analytic criterion.

Stochastic Control Theory

Let us return to the original problem (2), and ask what have we assumed? We've tacitly assumed the following: a) we know the initial state precisely; when we say it's c, we mean it's c; b) cause and effect are exact if we know the initial state and we know the initial control variable; c) the control is precise. We know the subsequent state exactly and we also know exactly what will happen when we exert a control v_n. In other words, we choose a value v, and we assume that's exactly what goes into the system; d) Finally, we have assumed that the criterion function is precise. Of course, in any actual physical system, none of these holds. We never are able to measure the state accurately, it's a matter of percentage error; we never can take account of all the different causes. Most of the time we don't even know many

many of them, and we don't know the effect of those we know precisely. Also, we never really know what happens when we choose a control variable. You might want a rocket to burn up at half maximum rate; it may be burning at half maximum rate, or it may not be burning at half maximum rate.

We have ignorances in each one of these four areas. The standard way of getting around ignorance is to assume that we can replace an unknown effect by random variables. And what we assume is that it's perfectly permissible then to introduce a random variable and take averages. This is an assumption. I keep insisting that you have to be very careful and make up your mind whether you want to take this very seriously. I say this because, unfortunately, in so much of the work that's done this is never mentioned. People assume that all this is on a completely rigorous basis -- not only rigorous mathematically, but rigorous scientifically. Whereas, the major problem is always to make sure that the physical situation is a good fit to the axioms that you use mathematically. This is part of the use of mathematics in science. You have to worry about this. Too many people do not worry about this.

Let's take the very simplest situation. Let's assume that we're now considering stochastic control, that the state at time $n + 1$ is a function of the state at time n, the control at time n, and a random influence at time n.

$$u_{n+1} = g(u_n, v_n, r_n), \quad u_0 = c \tag{7}$$

Let's simplify existence and assume that everything has been scaled down to the point where r_n at each stage is ± 1 with probability p, $1 - p$, respectively. This is perhaps the simplest kind of random variable and it illustrates all the complexities very well. Our cri-

terion function is

$$R_N = K(u_N) + \sum_{k=0}^{N-1} h(u_k, v_k) \quad . \tag{8}$$

We've assumed that we know what our criterion is; our criterion may have random effects in it too. Let's assume here that it doesn't.

The first problem we face, interestingly enough, is what we mean by a control process when we have uncertainty. There are two parts to the trick of probability. The first part is to say that we replace ignorance by knowledge. We don't exactly know what u_{n+1} is going to be. If we know u_n and V_n, the first part of the trick is that we circumvent that by introducing a random variable which can be +1 with probability p, and -1 with probability 1-p. So instead of complete ignorance, we say we'll try the one value or the other, and on the average, we know what the values are going to be.

The second question is, given this situation, how do you evaluate the outcome? The second part of the trick is to use expected values. In other words, we're going to use some average outcome, and we're going to evaluate the performance of the system in terms of the average outcome. As we do in baseball, we don't expect the pitcher to win every game or the team to win every game, or a batter to make a hit every time he goes to the plate, but we do evaluate their performance in terms of batting averages, team averages, etc. This standard technique is difficult to justify in many cases. Monday morning quarterbacks aren't interested in football coaches who say, well, my expected performance would have been excellent; we just had a few misfortunes, a few touchdowns here and a few touchdowns there.

If you're only interested in doing something once, such as surviving an H-bomb attack, you're not interested if somebody says your probability of survival is 0.9, because you worry a little bit about the remaining 0.1. In other words, one has to take this

application of probability theories with a grain of salt and make up your mind when you're interested in expected values and when you're not. It's clear that if you're going to do the same thing over and over and over again then the expected value means something; you have some general theorems in probability theory that tell you that if the average value of a random variable is p, over a long sequence of runs you're going to see approximately p of those values. Once again without worrying about whether one can use that in any particular situation, we're going to think in terms of expected values. We have to because nobody knows any better way for handling probability. Of course, the expected values need not be just the first moment. It can be the expected value of some function of r_n. Thus, we can also handle problems such as --- what is the probability that r_n exceeds a certain value? This is also an expected value. It's an expected value of a function of r_n. But, there is no way around dealing with expected values once you introduce probability theory.

So we agree to two things. We've introduced the idea of random influences and we're going to deal with expected values. To simplify life, let's say that we're going to minimize

$$\underset{[v]}{\text{Min}} \; \underset{[r]}{\text{Exp}} \; R_N \quad . \tag{9}$$

This sounds like a sensible problem. The point is, the problem is not defined as yet, and thus is meaningless. This is the amusing point. In the deterministic case I could set down the function R_n. I want to minimize it and not specify that I was talking about feedback control. The problem is well defined. As soon as we get into stochastic processes, you have to make the rules quite precise before you know what the problem is.

Let me point out two ways in which we can proceed in this case. To minimize the expected value of the R_N over the v's is not a precise problem. Let me point out two distinct types of problems that we could think of. The first I'll call non-sequential. Here we would choose $v_0, v_1, \ldots$ ahead of time; for any such set of values, compute

$$\underset{[r]}{\text{Exp}} \; R_N$$

and then minimize over the v_i. In other words I say one way of carrying on a control process of this type, is to say I'm going to pick my numbers $v_0, v_1, \ldots, v_{n-1}$ ahead of time. For each one of these choices I now compute the expected value of R_N over the r's. What I have left is a function of the v's. I now minimize in the usual way.

This is a meaningful engineering process. It corresponds to a situation in which you know that the system is going to operate in a certain way, you know the initial state, but you have no way of observing the system, once it gets started. If you have no way of observing the system, once it gets started, obviously, you cannot use a sequential process. So this is a meaningful process, in those situations in which you can obtain no information as to what the actual state of this system is, once you have started the control process.

Fortunately, in most cases, we can observe the state of the system. Then we proceed in the following way: Choose v_0, observe u_1, choose v_1, observe $u_2, \ldots$. This is a sequential or feedback control process.

Which one of these will yield a smaller value of the minimum? Obviously, the latter will, because nonsequential control is a subclass of this type. You can always exert nonsequential control.

In feedback control we have more information and we can expect to do better. But, as I say, unless I tell you what the rules of the game are, then, (9) is a meaningless problem. I have to tell you what inclination you're allowed at each stage.

This brings in a very interesting idea which we'll discuss later. That is the fact that as soon as you get to stochastic and adaptive processes, then the information pattern becomes important. What information do you have about a system at a particular time? This is something which doesn't enter at all into the deterministic case, because you tacitly assume that you know exactly what the behavior of the system is going to be. From a mathematical point of view, this means that stochastic and adaptive processes are infinitely more interesting. There are many more variations. I'll discuss some of them. We want now to consider (9). If I now say minimize over v, the feedback control,

$$(10) \qquad \underset{\substack{[v] \\ F.C.}}{\text{Min}} \quad \text{Exp} \quad R_N$$

I have a perfectly well defined process. Notice that as far as the sophistication of the problem is concerned, (9) is a lower level problem; it is a problem in calculus. We have to choose n-1 variables. (10) is a much more sophisticated problem because it is a problem involving the choice of n functions. We have to choose a policy. We say that once you have observed u_1, what should v_1 be as a function of u_1? u_1 will automatically be a range of values now as a stochastic variable; v_1 will be a function. In other words, in (10) we must make a choice of N policies. In solving the feedback control problem we have to choose a point in N dimensional space. However, although (9) is a much more elementary problem, (10) is a much easier problem to tackle. And the easier problem is the more important scientifically. There's a moral attached to this: complexity does not necessarily mean scientific importance. And very often, complexity and obscurity are just smoke screens to disguise the fact that there's very little there of scientific or intellectual interest.

Let us consider the feedback control problem in more detail. We proceed in exactly the same way as before. We say, clearly the minimum value over a feedback control process of the expected value over r depends on the initial value.

$$\underset{\text{F.C.}}{\underset{[v]}{\text{Min}}} \underset{[r]}{\text{Exp}}\, R_N = f_N(c) \tag{11}$$

It is a function of the number of stages and the initial state. Let us determine the corresponding recurrence relation. If we pick v_0, we incur immediate cost of $h(c,v_0)$. As a result of v_0 we're going to be in a state $g(c,v_0,r_0)$. Regardless of what state we're in, since r_0 is a random variable, it will be in one of two states. We're going to use the optimal continuation. We take the average value over the optimal continuations, of course. In this case the average value is just

$$f_N(c) = \underset{[v_0]}{\text{Min}} \left[h(c,v_0) + \underset{r_0}{\text{Exp}}\, f_{N-1}(g(c,v_0 r_0)) \right] \quad N \geqslant 1 \tag{12}$$

$$f_0(c) = \underset{[v_0]}{\text{Min}} \left[h(c,v_0) + \underset{r_0}{\text{Exp}}\, k\,(g(c,v_0,r_0)) \right].$$

These are our fundamental recurrence relations.

It requires a little bit of practice to juggle the minimization and the expected value in the right order; you have to think it out. In fact, you must read it backwards. First I choose v_0 which means I'm going to be in state $g(c,v_0,r_0)$. Whatever state I'm in I'm going to use an optimal policy from that point on. So, the return from any new state is going to be f_{N-1} of that. I don't know which one of these I'm going to have, but I average over the possibilities. f_0 is then my return. It's an average return. I now juggle the cost of the initial decision versus the cost of the remaining decisions. You have to play with these things for a while before you get confidence, and then it's very easy to interchange things and say, why don't you take the average value inside, etc. But we have to think of the process. As I pointed out yesterday, a certain amount of thought is necessary. This is good for one, it shouldn't be over done,

but a little bit never hurts. Let's take a specific example which may clarify things. Consider the simple linear situation

$$u_{n+1} = au_n + v_n . \tag{13}$$

And suppose my problem was

$$\underset{\substack{[v] \\ F.C.}}{\text{Min}} \; \underset{r}{\text{Exp}} \sum_{k=0}^{N} \left[(u_k - b)^2 + \lambda \sum_{k=0}^{N-1} v_k^2 \right] = f_N(c). \tag{14}$$

Then

$$f_N(c) = \min_v \left[(c-b)^2 + \lambda v^2 + p f_N(ac+v+1) + (1-p) f_N(ac+v-1) \right], \tag{15}$$

and

$$f_1(c) + \min_{v_0} \left[(c-b)^2 + \lambda v_0^2 + p \left[ac+v_0+1-b \right]^2 + (1-p) \left[ac+v_0-1-b \right]^2 \right], \tag{16}$$

where we have assumed that the function is just another constant.

For those of you who would like to do a little algebra and elementary calculus, let me pose the following problem:

Take the deterministic case first where

(17) $$u_{n+1} = au_n + v_n$$

with $$\min_{[v]} \left[\sum_{k=0}^{N} (u_k - b)^2 + \lambda \sum_{k=0}^{N-1} v_k^2 \right] = f_N(c).$$

Prove that $f_N(c)$ is a quadratic function of c---in other words, it has the form

(18) $$f_N(c) = \alpha_N + \beta_N c + \gamma_N c^2 .$$

Using the functional equation, derive recurrence relations for α_N, β_N, γ_N and show that optimal control is linear, i.e.

(19) $$v_n = \delta_n u_n + \epsilon_n .$$

This problem is discussed in Applied Dynamic Programming, referred to earlier. Do the same for the stochastic case. If you work through the details of this you will have some feel as to how these techniques can be used.

From the conceptual analytic point of view, the above approach gives us a uniform method for the treatment of both deterministic and stochastic control processes. We see that conceptually there's now no difference at all between the two; the concept of an optimal policy is exactly the same; each $v_k = v_k(c)$. An optimal policy now is, what control do you exert in terms of where you are? When we get into stochastic control process, you will see that we're talking in terms of feedback control. This is the way the solution has to be. Notice that the two techniques, the two approaches which were identical for deterministic processes, namely, either choose the control values all at once initially, or choose them sequentially, are quite different at the present time, and they represent quite different physical situations. As I pointed out, the a priori choice

is what's forced upon you when you have no way of determining what the actual state of this system is once you've proceeded with the control. On the other hand, the feedback control method depends upon the fact that you know the state of the system at each time.

A very interesting type of stochastic control process is one which we call interruptive control. Suppose we're controlling system (17), and suppose it really represents the behavior of a satellite. Assume that the communication link breaks due to interference at a certain time. You don't get a reading of the state variables. And now suppose that you know a priori that this is going to occur with a certain fixed probability. How do you control the system under such circumstances? We call this an interruption of the control process, and you see what I mean when I talk about the richness of stochastic control as opposed to deterministic control, because now you can take all the possible calculus of variations, all possible control theories, and systematically say, "Suppose I only know this with a certain probability. Suppose my information about this is lost or destroyed or interrupted. Then what do I do?" Each one of these problems is completely meaningful as far as an engineering process is concerned. Those of you who are further interested in stochastic control processes, I suggest that you look at Adaptive Control Processes: A Guided Tour. There's a much more detailed discussion there, and many references.

Speaking in inprecise terms mainly because there are no precise terms, an adaptive process is a process in which you have to learn about the system as you go along. We could really use the word learning process, but the psychologists have preempted that word. How does an adaptive process arise? I mentioned previously that feedback control or the feedback process is one of the most fundamental processes across the scientific board. More and more in the fields of biology and psychology, for example, people realize that there are all sorts of processes that formerly have been looked at in a rather mystical way but are just very simple examples of the feedback concept. Learning, for example, is one of these.

Let's consider the following quite precise process. Suppose I give you the stocastic process

$$u_{n+1} = g(u_n, v_n, r_n) \quad , \quad u_0 = c \tag{20}$$

with $r_n = \pm 1$ with probability p and 1-p, respectively.

I can set up the criterion function, and after I've done all that I say that, incidentally, I don't know the value of p. I give you (20), a very precise formulation, and then I say, as a postscript: p.s., p is not known. Now, of course, a strong tendency is to say, I don't admit that as a mathematical problem. I just refuse to deal with problems of that difficulty. But unfortunately in many applications, in engineering, in economics, in biology, this is the situation you face. As a matter of fact, this is a very simplified version of the situation you face. I want to talk about the more realistic aspects as we go along. You have a perfectly precise formulation a' la classical mathematics, and then it turns out that you don't know certain parameters or certain functions.

If you're a mathematician, you can just say "improperly posed," and you throw it away. If you're an engineer who has to construct a satellite or a spacecraft which, let's say, has to steer through an unknown atmosphere or will go out into regions of space where one doesn't know certain constants as well as one would like to know them, then you're faced with this problem; somebody says, I want to go to Mars or Venus; it's not for you to question why; just accept it. You look at the astronomical tables and it turns out that certain values of the parameters aren't know. What do you do?

You would say the following: if I'm going from, say, Terra to Mars and if I know how to get out just so far, of course I have many different paths. Let's assume that when I'm out in space I have instruments aboard my spacecraft which will enable me to take measurements much more accurately out there than from here, so that I can determine those unknown parameters. In other words, I have a control process in which I'm going to have to steer the ship. Also at certain stages I have inputs of new data.I must learn and do at the same time. This, of course, is a typical situation in life.

The second example is a person who is in charge of an Air Force depot. Suppose you had to store spare parts for a new plane. You cannot follow the obvious policy of saying let's order several thousand of these and several million of these, etc. That's rather expensive, and 5 or 10 years later, someone is going to look over the situation when you have 95% of this left over and 85% of that, and say, this is too expensive. The question is, how many spare parts do you order? The number of spare parts should depend upon the demand. If you're in a well established industry, you have a nice distribution of demand which says expected demand is so large and there is an expected probability of lower demands or excessive demands. The exercise in probability theory is to determine how many you should order so as to minimize your expected cost.

If this were a multi-stage process, one could use dynamic programming and get into what's called inventory problems.

Suppose you have a new item. You don't have any previous demand curve. You have to stock a certain number of parts, observe what happens over the first month or the first year and on the basis of that improve your estimate of the demand curve. You keep on going in that way. This again is, of course, a typical situaion in industry. As technology is changing tremendously rapidly, you're constantly in a new situation where you don't have experience to guide you as to what your probability distributions are, but you have to learn as you go along.

Another example of this is in connection with our missiles. Missiles have the peculiar property, because of their electronic gear, of falling into disrepair while just sitting there doing nothing. This means that people have to go around and look at them from time to time. Ideally, if you want to make sure that the thing is working, you ought to look at it all the time. While you're looking at it, it is operative and workmen will demand at least double pay if there is the probability that the thing will be fired while they're watching it. So, if you're turning out new devices like that, you have to determine what your inspection policy should be, on the basis of starting to inspect it and seeing what happens as you go along. There is no well extablished curve of what the probability is that something goes wrong.

In a period of rapid acceleration of technology where we're using new devices all the time, the really significant processes are the adaptive processes. It's rather interesting that the mathematicians, and the engineers, and the people in economics and history, and the scientists and the people in biology have finally

accepted the fact that the basic problem has always been this. We really have been in the situation of the people in "The Emperor's New Clothes." For hundreds and hundreds of years, these people have always assumed, first of all, that the systems were perfectly deterministic. Then finally, with much bitterness, they've brought in stochastic systems, and finally they're beginning to admit that the real problems are adaptive. The real situation is that you never know as much as you want to know about a complex system, but that you learn about it in the process of using it.

The flexibility of the feedback concept is absolutely essential. You say that you keep yourself in readiness to change; what you're going to do is dependent on what happens. In the world of biology, you can cite many examples of organisms which survive because they do have the feedback potential. Other organisms perish because their policies were so rigid that they had no way of adapting to new circumstances. Talking about those lines, you know, it's important to give up the phrase about the dinosaurs being a very unadaptable group. One should remember the dinosaurs existed for about a hundred million years; we've existed for about 5,000,000 years at most, and certainly maybe in only conscious form for about 15,000 years. So before we sneer at the dinosaur for not being able to adapt we ought to at least get past the next 50 years.

Another very interesting example of an adpative process occurs in the field of medical diagnosis. Here is the situation: you go to a doctor; you say, I don't feel well. He performs a certain number of obvious tests, gets a certain number of reactions, and makes a certain number of prescriptions; they give you a shot of this or a shot of that, or some aspirin, etc. Then he waits and sees. At the end of a day or two, or when you complain next, he looks at the situation again. And so on.

One further example of an adaptive process is the use of new wonder drugs. Suppose you have a drug which has never been tested before. The problem that a doctor has is: Should he prescribe something he knows will work with a certain probability, or something which is relatively untested? It may not work at all - it may work very well. The experimental implementation of new drugs is a very interesting process, which is an adaptive process.

As a matter of fact, this problem was first thought of in about 1932. The person who worked on it was an expert in biological medical statistics and he discovered sequential analysis. He realized that if you're going to try a new drug rather than the standard, prosaic technique in which you take a hundred cases here and a hundred cases there, what you ought to do is take ten cases or twenty cases, one with the new drug, one with the control, and depending upon what happens, change the size as you go along. He discovered sequential analysis and then he decided that the mathematics were too complex so he devised Monte Carlo techniques, in order to test this. This was in 1932. Unfortunately, he was 14 or 15 years ahead of his time and so his work went completely unnoticed.

It's rather interesting to see that many of the basic problems of scientific life are adaptive processes. The challenge is, if we face that much uncertainty, do we have techniques for handling it? I pointed out previously that probability theory was a very ingenious way of circumventing the fact that if we're in state u_n and if we apply control v_n we don't know what the state is going to be. The simplest example is to take a coin and toss it. We don't know if it's going to fall heads or tails: we know it definitely will be one or the other, assuming the coin is thin enough, but we can't predict. So we get around this fundamental difficulty by introducing random variables.

I'd like to point out for those of you who have had only traditional courses in probability theory, with very little discussion of the philosophical and conceptional difficulties, that the best book on the subject written to date and probably the best book that ever will be written was written by Laplace, his Essay on Probability Theory. An English translation is published by Dover. The book contains lectures that he gave in 1799 in Paris with the constraint that these were to be public lectures. No mathematical symbols were to be used. It's very interesting to see him talk about the Gaussain distribution without any symbols at all. When he mentions pi he never uses the symbol for π, he says the ratio of the circumference to the diameter of a circle. But he mentioned and discussed such problems as how many people should be on a jury and all sorts of problems which people think are just modern operations research. They forget that there was the same interest in the application of mathematics to the problems of the world at the beginning of the 19th century and the end of the 18th. And one of the reasons why this didn't flourish is because people realized very precisely then that mathematics was quite limited. One has to be very careful before you apply it to economic, social, or political problems. I think we're beginning to find that out again today. But this is an excellent book, completely readable, very charming.

Adaptive Control

by

Dr. Richard E. Bellman

Adaptive Control

The concepts of adaptive control can best be brought out by making use of the problems that we have discussed earlier in stochastic control theory, where certain of the quantities which were well known before now are considered to be less well known or entirely unknown. This requires the introduction of new techniques of analysis.

Consider the simple linear problem discussed previously:

$$(1) \qquad u_{k+1} + au_k + v_k + r_k, \qquad u_0 = c.$$

The criterion function we take to be a simple quadratic, and the problem is

$$(2) \qquad \underset{\substack{[v] \\ F.C.}}{\text{Min}} \; \underset{r}{\text{Exp}} \left[\sum_{k=0}^{N} (u_k - b)^2 + \lambda \sum_{k=0}^{N-1} v_k^2 \right],$$

where the first term represents the cost of deviation from the desired state b, and the second term is the cost of control. r at the j-th stage is either +1 or -1, with probability p and 1-p, respectively. The problem is identical to that posed in stochastic control theory with the following exception: p is unknown!

In order to remove the ambiguity introduced by trying to find $\underset{r}{\text{Exp}}$, if r is unknown, and hence posing an intrinsically meaningless problem, we must now deal with the probability of a probability, a technique which is common in the field of statistics.

Let us assume:

1. That p has itself an a priori probability distribution dG(p),

2. That we will revise this a priori distribution function on the basis of outcome as the process unfolds by, for example, the Baye's estimation, which is the simplest method,

3. We will act as if the expected probability is the actual probability.

In processes of this type the information pattern plays an important role. In addition, we must consider an enlarged concept of the state of the system. We now must consider a) the physical state c, b) the sequence of values of r_i, i.e., $[1, 1, -1, -1, -1, \ldots]$. In fact, the order is not important in this case. But it is essential to observe how many +1's and how many -1's occur, say m+1's and n-1's. We will now interpret the state in the generalized sense:

$$P = P(c, m, n). \tag{3}$$

A typical estimate of the probability of r would then be

$$p = \frac{m+1}{m+n+2}. \tag{4}$$

By the Baye's estimation formula,

$$dG(p) = \frac{p^m(1-p)^n dp}{\int_0^1 p^m(1-p)^n dp}. \tag{5}$$

Hence the expected probability,

$$p_{mn} = \frac{\int_0^1 p^{m+1}(1-p)^n dp}{\int_0^1 p^m(1-p)^n dp}. \tag{6}$$

This changes an a priori estimate to an a posteriori estimate. Of course, this must converge to the true probability for p=1.

The functional equations for problem (1) with criterion function (2) follow directly by means of the methods discussed earlier.

$$(7) \qquad \underset{\substack{v \\ F.C.}}{\text{Min}} \; \underset{r}{\text{Exp}} \left[\sum_{k=0}^{N} (u_k - b)^2 + \lambda \sum_{k=0}^{N-1} v_k^2 \right] = f_N(c, n, m).$$

And it follows that

$$(8) \qquad f_N(c, n, m) = \underset{v_0}{\text{Min}} \left[(c-b)^2 + \lambda v_0^2 + p_{mn} f_{N-1}(ac+v_0+1, m+1, n) + (1-p_{mn}) f_{N-1}(ac+v_0-1, m, n+1) \right].$$

Among these higher level control processes, there is a hierarchy of uncertainty. Starting at the lowest level of uncertainty we have:

1. Problems in which the uncertainty has a known probability. These we have called stochastic control processes;

2. Problems in which the uncertainty has an unknown probability, but the probability has a known distribution function; and

3. Problems in which the uncertainty has an unknown probability with an unknown distribution function, but the distribution function belongs to a family of functions characterized by a fixed but unknown parameter, itself possessing a distribution function, e.g.

$$(9) \qquad dG(p) = \frac{p^{\alpha}(1-p)^{\beta}dp}{\int_0^1 p^{\alpha}(1-p)^{\beta}dp}.$$

Here $dG(p)$ represents the unknown distribution function of p, and the probability distribution function of α and β may be another function $dH(\alpha,\beta)$.

Analagous to the hierarchy referred to above is the level of intelligence of so-called intelligent digital computers, provided one defines the level of intelligence as the level of ability to make decisions. At the present state of the art, our digital computers are at Level 0, i.e., they can handle only strictly deterministic processes. From a state p they obtain, by a simple transformation, state $p_1=T(p)$, then $p_2=T(p_1) = T^2(p)$, etc., no more than the simple iteration processes referred to at the beginning of this lecture series. The next level, Level 1, is that of stochastic processes; Level 2 is occupied by simple adaptive processes; and Level 3 contains the complicated learning processes outlined above. There is much doubt that we can ever make computers that will achieve the higher levels. To do so, we would have to understand completely the process of human thought.

Concluding Remarks

Dynamic programming is a method of handling multi-stage decision processes. To make use of the techniques, one must be able to convert problems which traditionally have been handled by classical methods, or be able to recognize among the large class of classically unmanageable problems those problems that can be interpreted as multi-stage decision processes. In most mathematical problems,

one doesn't know a priori which is the proper mathematical method.

For example, the simple problem of finding the solution of

$$(10) \quad u'' - u = 0, \qquad u(0) = 1, \qquad u'(0) = 0$$

is entirely equivalent to

$$(11) \quad \min_{u(0)=1} \int_0^1 (u'^2 - u^2)\,dt.$$

The latter case is easily interpreted as a multi-stage decision process.

Another simple example is

$$(12) \quad \text{Max} \sum_{i=1}^{N} x_i \quad \text{over} \quad \sum_{i=1}^{N} x_i^2 = a$$

which, when properly interpreted, leads to the functional equation

$$(13) \quad f_N(a) = \max_{x_N} \left[x_N + f_{N-1}\,(a - x_N^2) \right]$$

which is solved by choosing x_N to maximize (13), then x_{N-1}, etc.

The concept of a state plays an important role in the theory. By means of a transformation T, the state $p \longrightarrow T(p)$, and then control q is easily conceived as a choice of transformation, i.e., the controlled transformation is $p \longrightarrow T(p, q)$. The criterion function usually is entirely

arbitrary, and can be chosen for convenience. The central question here, as in all of mathematical physics, is whether to choose, to represent some physical situation, a set of complicated equations and then solve by means of some approximate methods, or to represent the system approximately and solve the corresponding equations exactly. In control theory we approximate a complicated system by simple models and solve these in the best possible way.

ELLIPTIC MOTION

by

J.M.A. Danby

Chapter One: Introduction to the Mechanics of Celestial Bodies.

1.1 Newton's Law of Gravitation.

Celestial mechanics is, in the main, a branch of Newtonian mechanics, and the fundamental law is Newton's Law of gravitation. It is true that this law may not cover every contingency in cosmogony, but its inadequacies in celestial mechanics are small indeed. Also, in cases where the relevant arguments of general relativity have achieved explicity forms, the resulting arguments of general relativity have achieved explicit forms, the resulting modifications to motion governed by Newton's laws have been dealt with by established perturbation theories of celestial mechanics. (5)

Newton's law states that: "any two particles attract each other with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them." Let the particles ave masses m_1 and m_2, and position vectors $\underline{r}_1$ and $\underline{r}_2$, respectively. Then the force exerted by m_2 on m_1 can be written

$$- k^2 m_1 m_2 (\underline{r}_1 - \underline{r}_2) / | \underline{r}_1 - \underline{r}_2 |^3 .$$

(In these notes a vector is written with a line underneath. A unit vector is written with a "cap" above it, i.e., $\hat{r}$, and $| \underline{r} |$, or simply "r" stand for the modulus of $\underline{r}$. It is assumed that the reader is acquainted with elementary vector algebra and calculus; if not, see (2). k^2 is often written as "G", and the value in c.g.s. units is $6.67 \cdot 10^{-8}$. But this value would not be accurate enough for calculation, and normally special units must be chosen so that the constant is known much more accurately.

Consider a system of masses $m_1, m_2, \ldots, m_n$ at $\underline{r}_1, \underline{r}_2 \ldots, \underline{r}_n$. They will exert a total force on a mass m at $\underline{r}$ of amount

$$- k^2 m \sum_{i=1}^{n} m_i (\underline{r} - \underline{r}_i) / | \underline{r} - \underline{r}_i |^3 .$$

The particle m will experience some force wherever it is, and the n bodies are said to set up a "field of force." The strength of a field of force at a point $\underline{r}$ is the force exerted on a particle of unit mass placed at $\underline{r}$. Strictly, the word "force" in this context means "force per unit mass." The n bodies produce, therefore, a field of force

$$- k^2 \sum_{i=1}^{n} m_i (\underline{r} - \underline{r}_i) / | \underline{r} - \underline{r}_i |^3 . \qquad (1.1.1)$$

The three components of force in (1.1.1) can be derived from the gradient of the "force function"

$$U = k^2 \sum m_i / | \underline{r} - \underline{r}_i | . \qquad (1.1.2)$$

For instance, the x-component of the gradient of U (or gradU, or ∇U) is $\partial U / \partial x$. Since

$$| \underline{r} - \underline{r}_i | = (x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2 \ ^{-1/2} ,$$

then $$\partial | \underline{r} - \underline{r}_i | / \partial x = (x - x_i) \ (x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2 \ ^{-1/2} ,$$

and $$\partial U / \partial x = - (x-x_i) \ (x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2 \ ^{-3/2} .$$

The force function is the negative of the work that would be done in assembling the system of n bodies from a state of infinite diffusion. As the words are normally used, it is minus the potential; but this convention is not universal, and I shall use only force functions here.

The transition from particles to solid bodies is accomplished by integration. Consider the force function of a uniform, thin spherical shell at a point O outside the shell. Let the shall have center C, radius a, thickness da, and density ρ; and let $OC = r$. If P is a point on the shell, let the angle $\angle OCP = \theta$. Divide the shell into thin rings perpendicular to OC and defined by θ lying within the limits θ and $\theta + d\theta$. The radius of a ring is $a\sin\theta$,

and its mass is

$$\rho 2\pi a \sin\theta \; a \, d\theta \, da.$$

Any element of the ring is at

the distance

$$(r^2 + a^2 - 2a\cos\theta)^{1/2}$$

from O, so that the force function

of the ring at O is

$$k^2 \rho 2\pi a^2 da \sin\theta \, d\theta \, (r^2 + a^2 - 2a\cos\theta)^{-1/2},$$

and the total force function due to the shell is

$$U = k^2 \rho 2\pi a^2 da \int_0^{\pi} \sin\theta d\theta \, (r^2 + a^2 - 2a\cos\theta)^{-1/2},$$

where the square root must always be positive. This can be integrated at once to give

$$U = \frac{1}{2} k^2 \, dm \left[\frac{1}{ra} (r^2 + a^2 - 2a\cos\theta)^{1/2} \right]_{\theta = 0}^{\pi}$$

$$= k^2 \, dm/r,$$

where $dm = 4\pi a^2 \rho \, da$ is the mass of the shell. This means that, so far as O is concerned, the shell could just as well have all its mass concen-

trated at C. This must also apply to a shell of finite thickness, since the result is not affected by integration over a, and it applies in particular to any solid body that is constructed in concentric spherical shells, provided we are outside it.

If, therefore, we have a system of n bodies, each having spherical symmetry, then they can be considered as particles generating a force function (1.1.2), provided they do not approach too close to each other. The mass of each body is considered to be concentrated at its center of gravity, and the coordinates of a body are the coordinates of its center of gravity.

Fortunately, in most problems of celestial mechanics the bodies can be assumed to be spheres. In the first place they are, in fact, nearly spherical, and in the second place the distances between the bodies are usually large compared with the dimensions of the bodies themselves. In the case of the motion of an artificial satellite the latter condition does not hold, and the oblateness of the Earth actually causes major perturbations in the motion.

Outside a gravitating body the force function must satisfy Laplace's equation,

$$\nabla^2 U = \partial^2 U/\partial x^2 + \partial^2 U/\partial y^2 + \partial^2 U/\partial z^2 = 0.$$

(This can be proved by differentiating equation (1.1.2); the summation is replaced by an integration.) It transpires that the force function of the body can normally be expanded in a power series in $1/r$, where

r is the distance from its center of mass; the coefficients are called spherical harmonics. If, as is often the case, the body has symmetry about an axis, the force function can be expressed as

$$\frac{Mk^2}{r}\left(1 - \frac{1}{r^2} J_2P_2 - \frac{1}{r^3} J_3P_3 - \ldots\right), \qquad (1.1.3)$$

where the P_i are Legendre polynomials (functions of the latitude) and the J_i are constants; if the body is nearly spherical, the latter becomes small quite rapidly. Now it would be possible to find the force function of such a body by integration, if we knew precisely how it was put together. Failing this knowledge, it is still possible to write down its force function directly, so far as all the variable quantities are concerned. The theory of the motion of an artificial satellite, without drag, can be constructed using the force function (1.1.3). Then, later, observations may furnish the values of the J_i. A lack of knowledge about the insides of a body is therefore no great hardship when its force function is required. (For more details, see Ref. 2, Chapter 4.)

1.2 Newton's Laws of Motion.

We are concerned with Newtonian mechanics, the basic assumptions of which are contained in Newton's laws of motion. These are:

1. Every particle continues in a state of rest or uniform motion in a straight line unless it is compelled by some external force to change that state.

2. The rate of change of the linear momentum of a particle is proportional to the force applied to the particle and takes place in the same direction as that force.

3. The mutual actions of any two bodies are always equal and oppositely directed.

A man who observes the motion of surrounding bodies that are not acted on by forces, and notes that they are not accelerated is entitled to feel that, for practical purposes, he is at rest with respect to some inertial system of reference. But if these bodies have any accelerations, then he is not (although he may invent forces such as centrifugal or Coriolis forces, to preserve the illusion). Certainly, no point fixed on the surface of the Earth could be the origin of an inertial system, although some sufficiently parochial experiments might give that impression. Motion observed (sic) by a non-rotating man at the center of the Earth would still show acceleration because of the action of the Sun, Moon, etc., on the Earth. Similarly, motion observed from the center of mass of the solar system should be affected by nearby stars, and the field of the galaxy (to say nothing about nearby galaxies): this is true in principle; but there is no known experiment to detect such effects, so that no purpose is served by considering acceleration with respect to the center of the galaxy, and so on. So we shall not worry about the practical difficulties of choosing an inertial reference system, and we are certainly not concerned here with the thornier difficulties as to whether such a system

can exist at all. We adopt the attitude that, given any problem in Newtonian mechanics, there exists an inertial system with respect to which the equations of motion can be written down; but no special assumption must be made about the whereabouts of the origin. Once the equations of motion have been set up, algebra will enable the origin to be transferred to this place or that. Also, inspection of some terms in the equations may result in their being rejected on account of their smallness. But the original equations must be written down without any assumptions being made about the origin, or the relative importance of different terms.

The measurement of "uniform motion" requires the use of a "uniformly flowing" time. The use of Universal Time (which is based on the rotation of the Earth) threw up accelerations of the Moon and planets that could not be explained by Newtonian mechanics, but which could result from non-uniform flowing of Universal Time. A suitable time has therefore been invented; this is Ephemeris Time. Its relation with Universal Time is given in the almanacs.

The second law can only be applied to motion observed with respect to an inertial reference system. If a particle of mass m is at $\underline{r}$ and the resultant of the forces acting on the particle is $\underline{F}$, then

$$\underline{F} = \frac{d}{dt}\left(m\frac{dr}{dt}\right). \qquad (1.2.1.)$$

Two important formulas follow from this. Firstly,

$$\underline{r}\times\underline{F} = \frac{d}{dt}\left(\underline{r}\times m\frac{dr}{dt}\right), \qquad (1.2.2.)$$

or "the moment of the external forces is equal to the rate of change of the angular momentum". Then, if $\underline{F}$ is the gradient of a force function U that does not contain the time explicitly, and if m is constant,

$$\frac{1}{2}\left(\frac{dr}{dt}\right)^2 - U = \text{constant}. \qquad (1.2.3)$$

(For, differentiating (1.2.3) with respect to the time, we have the scalar product of $d\underline{r}/dt$ and (1.2.1).) This is the energy integral.

The third law is obeyed by Newton's Law of gravitation, and is needed in a derivation of this law from Kepler's laws of planetary motion (quoted in Section 2.4).

Newton's laws apply directly to the motion of particles. If a body of finite extent is acted on by a system of forces, then the motion of its center of mass can be found by shifting the forces parallel to themselves so that their lines of action pass through the center of mass. The motion about the center of mass is considered, basically, through equation (1.2.2); subjects such as precession or physical libration fall under this heading; but they will not be considered here.

1.3 Equations of Motion.

Consider the motion of n particles with masses $m_1, m_2, \ldots, m_n$, which have position vectors $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_n$, with respect to an inertial reference system. The equation of motion of m_i is

$$m_i \boldsymbol{\rho}_i'' = -k^2 m_i \sum_{j=1}^{n} m_j \frac{\boldsymbol{\rho}_i - \boldsymbol{\rho}_j}{\rho_{ij}^3}, \qquad (j \neq i) \qquad (1.3.1)$$

where $\rho_{ij} = |\rho_i - \rho_j|$. A prime stands for differentiation with respect to the time. Adding the equations for all the particles, the forces cancel (from the algebra, or from Newton's third law) leaving

$$\sum_{i=1}^{n} m_i \rho_i'' = 0. \tag{1.3.2}$$

But $\sum m_i \rho_i$ is the position vector of the center of mass of the system, and this is not accelerated with respect to the original inertial system; therefore the center of mass could be the origin of an inertial reference system.

Multiply (1.3.1) vectorially by $\underline{r}_i x$, and add all n equations. The right hand sides again cancel, leaving

$$\sum_{i=1}^{n} m_i \underline{r}_i x \underline{r}_i'' = 0, \quad \text{or} \quad \sum_{i=1}^{n} m_i \underline{r}_i x \underline{r}_i' = \underline{h} \tag{1.3.3}$$

where $\underline{h}$ is a constant vector. The plane through the center of mass of the system and perpendicular to $\underline{h}$ is constant throughout the motion, and is called the "invariable plane" of the system.

The equations (1.3.1) can be written in the form

$$m_i \rho_i'' = \Delta_i U, \tag{1.3.4}$$

where, if ρ_i has components (ξ_i, η_i, ζ_i), ∇_i has components $\partial/\partial\xi_i, \partial/\partial\eta_i, \partial/\partial\zeta_i$, and where

$$U = k^2 \sum_{i<j=1}^{n} \frac{m_i m_j}{\rho_{ij}}. \tag{1.3.5}$$

We therefore have the energy integral for the whole system,

$$\frac{1}{2}\sum_{i=1}^{n} m_i \rho_i'^2 - U = \text{constant.} \qquad (1.3.6)$$

But no integral can be written down for an individual member of the system.

Suppose that one body, m_n, is considered to be dominant, either because of its relatively great mass, or because the motion in which we are interested takes place very close to it. Subtracting the equation of motion of m_n from that of m_i (after dividing by m_n and m_i, respectively) we find

$$\rho_i'' - \rho_n'' = -k^2 \sum_{\substack{j=1 \\ j\neq i}}^{n} m_j \frac{\rho_i - \rho_j}{\rho_{ij}^3} + k^2 \sum_{j=1}^{n-1} m_j \frac{\rho_n - \rho_j}{\rho_{nj}^3}$$

$$= -k^2 m_n \frac{\rho_i - \rho_n}{\rho_{in}^3} + k^2 m_i \frac{\rho_n - \rho_i}{\rho_{ni}^3} - k^2 \sum_{\substack{j=1 \\ j\neq i}}^{n-1} m_j \left[\frac{\rho_i - \rho_j}{\rho_{ij}^3} - \frac{\rho_n - \rho_j}{\rho_{nj}^3} \right].$$

Now let the position vector of m_i with respect to m_n be $\underline{r}_i$, so that $\underline{r}_i = \rho_i - \rho_n$. Then

$$\underline{r}_i'' + k^2(m_n + m_i)\frac{\underline{r}_i}{r_i^3} = k^2 \sum_{\substack{j=1 \\ j\neq i}}^{n-1} m_j \left(\frac{\underline{r}_j - \underline{r}_i}{r_{ij}^3} + \frac{\underline{r}_j}{r_j^3} \right). \qquad (1.3.7)$$

Further, if

$$R_{ij} = k^2 \left[\frac{1}{r_{ij}} - \frac{\underline{r}_i \cdot \underline{r}_j}{r_j^3} \right], \qquad (1.3.8)$$

then

$$\underline{r}_i'' + k^2(m_n + m_i)\frac{\underline{r}_i}{r_i^3} = \sum_{\substack{j=1 \\ j\neq i}}^{n-1} m_j \Delta_i R_{ij}. \qquad (1.3.9)$$

Now if all the masses except m_n and m_i were zero, the right hand sides of (1.3.6) or (1.3.8) would vanish, and the equations of motion would refer to the two-body problem; the solution of this is called Keplerian motion, and is described in the following chapter. It is frequently possible in celestial mechanics to find a dominant body, m_n, such that the terms on the right hand side of (1.3.7) are much smaller than $k^2(m_n + m_i)\underline{r}_i/r_i^3$. In this case the motion can be considered as Keplerian motion "perturbed" by the forces on the right hand side. This is why Keplerian motion is so important in celestial mechanics. The word "perturbation" normally implies a departure from Keplerian motion; the forces on the right hand side of (1.3.7) are "perturbing forces" and the R_{ij} are "perturbative functions".

The reference system in (1.3.7) is non-inertial. The terms on the right hand side include the "direct" attractions of the bodies on m_i, and the "indirect" attractions on m_n, the origin. In a practical application many of these terms might be found to be negligible; but it can happen that the direct attraction of a body is negligible, but the indirect attraction is not. Further modifications can be made by adjusting the origin, and the mass of the dominant body; for details, see Ref. 2, section 9.5.

Chapter Two: The Two-Body Problem.

2.1 Properties of Conics.

Any orbit in the two-body problem is a conic, and before discussing the solution we shall briefly review the relevant properties of conics.

The polar equation of a conic can be written as

$$p/r = 1 + e\cos f, \qquad (2.1.1)$$

where the origin is at a focus of the conic, and f is the polar angle, measured from the major axis. [Equation (2.1.1) follows from the "focus-directrix" definition of a conic; i.e., that it is the locus of a point such that the ratio of its distance from a fixed point (a focus) to its distance from a fixed line (a directrix) is constant, the value of the constant being equal to e.] e is the eccentricity; if e = 0 the conic is a circle; if e is less than one, it is an ellipse, which is bounded; if e = 1 it is a parabola; if e is greater than one it is a hyperbola.

Let an ellipse have center C, foci S, S', major axis AA', and minor axis BB'. The following relations are useful, and should be memorized:

$CA = CA' = a,$

$CB = CB' = b,$

$SA = q = a(1-e),$

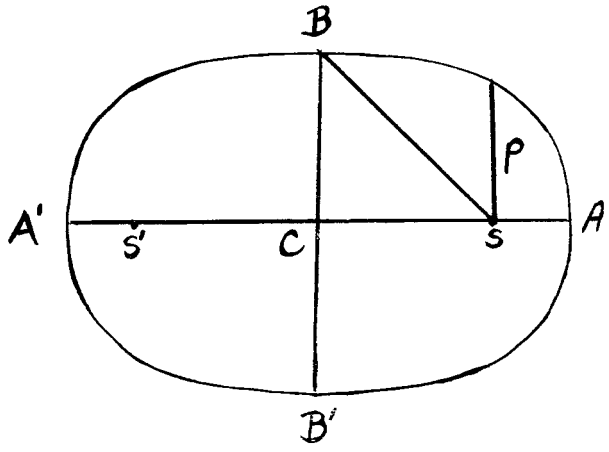
$SA' = q' = a(1+e),$

$CS = CS' = ae,$

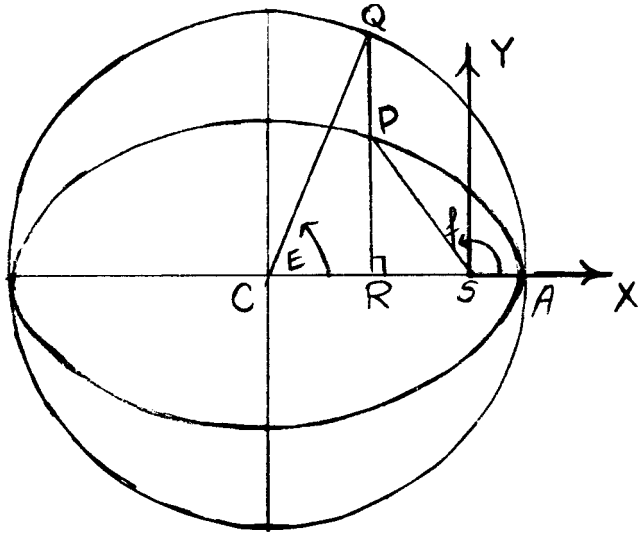
$p = a(1 - e^2),$

$b^2 = a^2(1 - e^2),$

$SB = a.$



The ellipse can be obtained by the orthogonal projection of a circle. Let Q be a point on the circumference of the circle, and P the corresponding point on the ellipse, and let QP cut the major axis at R. Then $PR/QR = b/a$. Further, let $\angle ACQ = E$ (the "eccentric anomaly"). With origin at S, let the X-axis point along SA, and the Y-axis point along the latus rectum, as shown. This reference system will be called the "orbital reference system". Then the coordinates of P can be written:



$$X = a(\cos E - e) = r\cos f, \qquad Y = b\sin E = r\sin f. \qquad (2.1.2)$$

The area of the ellipse is πab. We also have

$$r = \sqrt{X^2 + Y^2} = a(1 - e\cos E). \qquad (2.1.3)$$

Formulas for the parabola can be obtained from those for the ellipse by (carefully) letting $a \to \infty$ and $e \to 1$. It is safest first to eliminate a or e using $q = a(1 - e)$, since q remains finite. Suitable modifications to cover hyperbolic motion will be given in section 2.4.

2.2 The Solution of the Orbit.

Consider two particles of mass m_1 and m_2. Let the position vector of m_2 with respect to m_1 be $\underline{r}$. From (1.3.7) we see that the equation of motion of m_2 is

$$\underline{r}'' + k^2(m_1 + m_2)\underline{r}/r^3 = 0. \qquad (2.2.1)$$

If the origin was at the center of mass of the two bodies, the reference system (non-rotating) would be inertial. Then if the masses were at $\underline{r}_1$ and $\underline{r}_2$, the equation of motion of m_2 would be

$$m_2\underline{r}_2'' = -k^2 m_1 m_2 \underline{r}/r^3$$

or, since $\underline{r}_2 = \left[m_1/(m_1 + m_2)\right] \underline{r}$,

$$\underline{r}_2'' = -k^2\left[m_1^3/(m_1 + m_2)^2\right]\underline{r}_2/r_2^3. \qquad (2.2.2)$$

Equations (2.2.1) and (2.2.2) are of the form

$$\underline{r}'' = -\mu\underline{r}/r^3 \qquad (2.2.3)$$

but with different values of μ.

Equation (2.2.3) requires six constants of integration for its solution. Taking $\underline{r}$x(2.2.3), we find $\underline{r}\underline{x}\underline{r}'' = 0$, so that

$$\underline{r}\underline{x}\underline{r}' = \underline{h}, \text{ a constant.} \qquad (2.2.4)$$

$\underline{h}$ supplies three arbitrary constants. From (2.2.4), $\underline{r}.\underline{h} = 0$, which is the equation of a plane through the origin. The motion must take place in this plane; $\underline{h}$ determines its orientation, as well as the magnitude of the angular momentum. Now take $\underline{h}$x(2.2.3), and use (2.2.4). We find

$$\begin{aligned}
\underline{h}\underline{x}\underline{r}'' &= -\frac{\mu}{r^3}(\underline{r}\underline{x}\underline{r}')\underline{x}\underline{r} \\
&= -\frac{\mu}{r^3}\left[r^2\underline{r}' - (\underline{r}.\underline{r}')\underline{r}\right] \\
&= -\frac{\mu}{r^3}\left[r^2\underline{r}' - (rr')\underline{r}\right] \\
&= -\mu\left[\underline{r}'/r - \underline{r}r'/r^2\right] \\
&= -\mu\frac{d}{dt}(\underline{r}/r) \\
&= -\mu d\hat{r}/dt.
\end{aligned}$$

[$\underline{r}'$ is the velocity vector; r' is the rate of change of the scalar r. Differentiating $\underline{r}^2 = r^2$, we find $\underline{r}.\underline{r}' = rr'$; a useful relation.]

Integrating, we obtain

$$\underline{h}\underline{x}\underline{r}' = -\mu\hat{r} - \underline{P}, \qquad (2.2.4)$$

where $\underline{P}$ is an arbitrary vector; but since it is perpendicular to $\underline{h}$,

it only contains two arbitrary constants. The remaining constant of the motion will be considered in the following section. Taking $\underline{r}.(2.2.4)$ we obtain

$$\underline{r}.(\underline{h}\times\underline{r}') = -\mu r - \underline{P}.\underline{r}$$

or

$$-\underline{h}.(\underline{r}\times\underline{r}') = -\mu r - \underline{P}.\underline{r}$$

or

$$h^2 = \mu r + \underline{P}.\underline{r}$$

or

$$\frac{h^2/\mu}{r} = 1 + (\underline{P}/\mu).\hat{r}.$$

This is the same as equation (2.1.1). We have $h^2/\mu = p$, the vector $\underline{P}$ points along the major axis toward pericentron, and $P = \mu e$. The angle f is called the "true anomaly". If e is greater than one, only the branch of the hyperbola that is concave toward the origin can be described in the motion.

2.3 The Orbit in Time.

The vector $\underline{r}'$ has components r' along $\hat{r}$ and rf' perpendicular to it; therefore the modulus of $\underline{r}\times\underline{r}'$ is r^2f', which is twice the rate of change of the area swept out by the radius vector. From (2.2.4) we have

$$r^2 df/dt = h. \qquad (2.3.1)$$

The integration of this equation supplies the final constant of integration. Substituting for r from (2.1.1) we get a simple integral; but except when e = 1, it is convenient to introduce an intermediate

angle, the eccentric anomaly.

Assume the motion to be elliptic. Differentiating (2.1.3) we find

$$r' = ae\sin E\, E'.$$

And differentiating $r\cos f = a(\cos E - e)$, (from (2.1.2)),

$$r'\cos f - r\sin f\, f' = -a\sin E\, E'.$$

Eliminating r' and f' from these two equations and (2.3.1) we find

$$h\sin f = a\sin E\, E'(1 + e\cos f)r.$$

Now using the relation

$$h^2 = \mu p = \mu a(1 - e^2),$$

and the formulas (2.1.2) and (2.1.3) to eliminate f and r, we find

$$\sqrt{\mu/a^3} = (1 - e\cos E)E',$$

which can be integrated to give

$$\sqrt{\mu/a^3}\,(t - T) = E - e\sin E,$$

where T is a constant of integration; it is equal to the time when E = 0, or when the body is at pericentron. This is Kepler's equation.

By letting the eccentric anomaly go from 0 to 2π, we get the time for a complete revolution, or the period of the motion, which is

$$P = 2\pi\sqrt{a^3/\mu}. \qquad (2.3.2)$$

The "mean motion", n, is defined by

$$n = 2\pi/P, \text{ so that } n^2a^3 = \mu. \tag{2.3.3}$$

The angle

$$M = n(t - T) \tag{2.3.4}$$

is defined as the "mean anomaly". So Kepler's equation can be written as

$$M = E - e\sin E. \tag{2.3.5}$$

Normally we are given the time, and want to calculate E. That there is a unique solution can be seen from the fact that the right hand side of (2.3.5) is monatonic increasing with E (for its differential coefficient with respect to E is (1 - ecosE), which is always positive). One of the best ways to find E is to use Newton's method. If E_0 is a good guess, and E is correct, let

$$\Delta E = E - E_0,$$

and

$$\Delta M = M - M_0 = M - E_0 + e\sin E_0.$$

Then if $(\Delta E)^2$ is neglected,

$$\Delta E = \Delta M/(1 - e\cos E_0).$$

Because of the approximation, this correction is not exact, and the process will have to be repeated until ΔM becomes less than some small pre-assigned value. This process converges best when e is small, when a good first guess is

$$E_0 = M + e\sin M,$$

(although the series for E in terms of M and powers of e, given in the following chapter, can be truncated later if desired). For more details, and for a discussion of the situation when e is nearly equal to one, see Ref. 3.

2.4 Miscellaneous Properties.

Kepler's three laws of planetary motion are:

1. The orbit of each planet is an ellipse, with the Sun at one of its foci. (Actually "Keplerian motion" is often now taken to include parabolic and hyperbolic motion, so that "conic" might replace "ellipse".)
2. Each planet revolves so that the line joining it to the Sun sweeps out equal areas in equal intervals of time. (Therefore the acceleration of the planet is directed toward the Sun, and so also is the force acting on the planet. From this law, and the first, Newton's law of gravitation can be deduced.)
3. The squares of the periods of any two planets are in the same proportion as the cubes of their mean distances from the Sun. This law should be modified so that $P^2(m_1 + m_2)/a^3$ is a constant for any two bodies, where a is the semimajor axis of the relative orbit, P is the period and m_1 and m_2 are the masses of the bodies. The law can be used to find the mass of a planet that has a satellite.

Many important formulas for elliptic motion have been given already.

A notable omission is the energy integral,

$$\underline{r}'^2 = \mu(2/r - 1/a). \tag{2.4.1}$$

The parabolic velocity, or velocity of escape is found by putting $1/a = 0$. The circular velocity is found by putting $r = a$.

When changing from E to f or f to E, the following formulas are useful:

$$\begin{aligned} \cos f &= (\cos E - e)/(1 - e\cos E), \qquad \sin f = \sqrt{1-e^2}\,\sin E/(1 - e\cos E), \\ \cos E &= (e + \cos f)/(1 + e\cos f), \qquad \sin E = \sqrt{1-e^2}\,\sin f/(1 + e\cos f). \end{aligned} \tag{2.4.2}$$

Using the relation $\tan^2(f/2) = (1 - \cos f)/(1 + \cos f)$, it is easy to verify that

$$\tan(f/2) = \sqrt{\frac{1 + e}{1 - e}}\tan(E/2). \tag{2.4.3}$$

When using these formulas, it should be remembered that $f/2$ and $E/2$ always lie in the same quadrant. If we write $e = \sin\phi (0 \leqslant \phi < \pi/2)$, as is commonly done, then

$$\tan(f/2) = \tan(\pi/4 + \phi/2)\tan(E/2).$$

From Kepler's equation, and (2.1.3) we have

$$E' = na/r. \tag{2.4.4}$$

Also we have

$$r' = na^2 e\sin E/r = (e\mu/h)\sin f. \tag{2.4.5}$$

Formulas (2.1.2) are important. Differentiating them, we find

$$\left.\begin{aligned} X' &= -na^2\sin E/r, \\ Y' &= na^2\sqrt{1-e^2}\cos E/r. \end{aligned}\right] \qquad (2.4.6)$$

In parabolic motion let q be the pericentron distance, then the equation of the orbit is

$$r = q\sec^2(f/2).$$

An equation involving the time is

$$\tfrac{1}{3}\tan^3(f/2) + \tan(f/2) = \sqrt{\mu/2q^3}\,(t - T).$$

Formulas for hyperbolic motion can be derived from those for elliptic motion as follows. Assume a to be negative for a hyperbola. If $i^2 = -1$, replace E by iF (so that cosE becomes coshF and sinE becomes isinhF), replace n by $-i\nu$, where $\nu^2 a^3 = \mu$, and ν is positive, and replace $\sqrt{1-e^2}$ by $i\sqrt{e^2-1}$.

Many important formulas have been omitted here. The reader should consult, in particular, Ref. 3.

2.5 The Orbit in Space.

An orbit is defined by six constants, and these require some kind of reference system. The celestial equator or ecliptic are often used as reference planes, with the direction of the vernal equinox defining an axis. Neither of these planes is fixed, and it is necessary

to use their mean positions for some definite epoch.

A suitable set of constants would be the components of position and velocity, $\underline{r}_0$, $\underline{r}_0'$, at some time t_0; it is possible to calculate from these the position $\underline{r}$ at any time t (formulas for the calculation of the velocity are easily deduced and will not be given here). Since the motion takes place in a plane, it must be possible to resolve $\underline{r}$ along the directions of $\underline{r}_0$ and $\underline{r}_0'$. So we can write

$$\underline{r} = f\underline{r}_0 + g\underline{r}_0',$$

where f and g are scalar functions of t_0 and t and the initial conditions. From (2.5.1) we find

$$f\underline{h} = \underline{r}x\underline{r}_0' \quad \text{and} \quad g\underline{h} = \underline{r}_0 x\underline{r}.$$

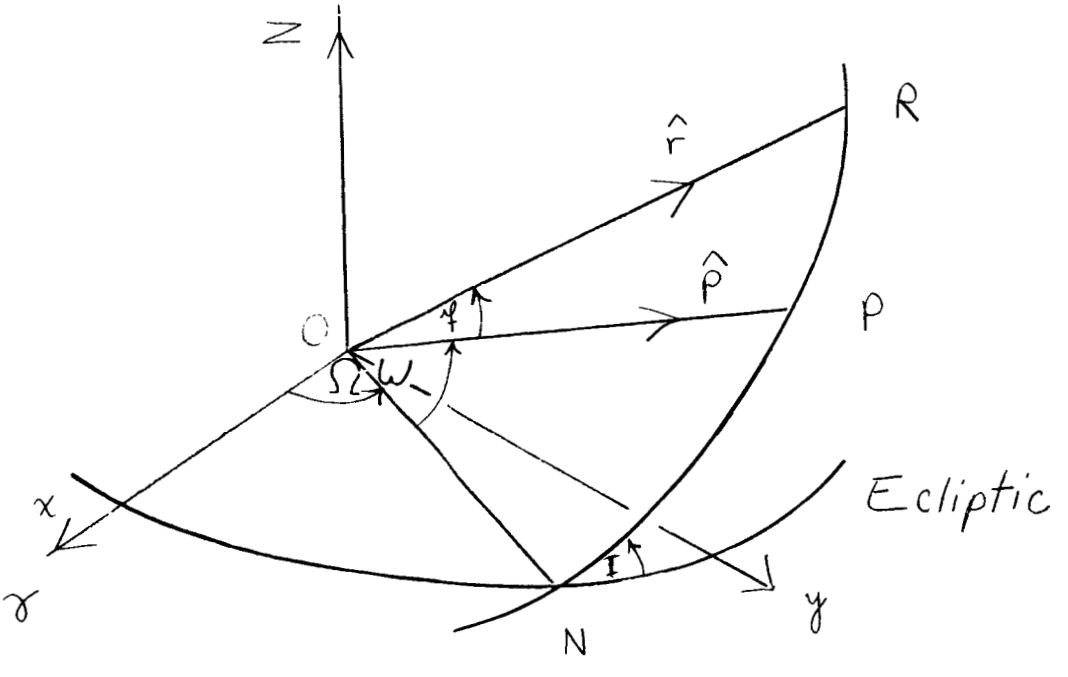
These are vector equations, independent of the reference system. So f and g can be evaluated by substituting components referred to the "orbital reference system" defined in section 2.1. The components are given by formulas (2.1.2) and (2.4.6). After substitution and some simplification, we find

$$\left.\begin{aligned} f &= \frac{a}{r}\left[\cos(E - E_0) - e\cos E_0\right], \\ g &= \frac{1}{n}\left[\sin(E - E_0) - e(\sin E - \sin E_0)\right]. \end{aligned}\right\} \qquad (2.5.2)$$

Before using (2.5.1) and (2.5.2) to calculate $\underline{r}$, it would be necessary to calculate a, e, E_0, and E. We are given $\underline{r}_0$ and $\underline{r}_0'$. (2.4.1) will give a. Then (2.1.3) and (2.4.5) will give $e\cos E_0$ and

$e\sin E_0$, from which e and E_0 can be found. (In using (2.4.5), remember that $r_0 r_0' = \underline{r}_0 . \underline{r}_0'$.) Finally, Kepler's equation can be used to find E.

In the formulas above a, e, and E_0 are introduced as intermediate elements; but they help to give a picture of the shape and size of the orbit, and the initial whereabouts in the orbit, that $\underline{r}_0$ and $\underline{r}_0'$ completely fail to do. It is more usual to use six constants, each of which has an easily visualized geometrical meaning; these are the "geometrical elements" of the orbit. a and e are two possible elements, and a third is a time of pericentron passage, T, or any number, such as the mean anomaly at some time, that enables the position in the ellipse to be found at any time. The description of the orientation of the orbit in space requires three angles, illustrated below.



In the figure the fundamental plane is the ecliptic (it could equally well be the celestial equator, if preferred), the Sun is at O, Ox points toward the vernal equinox and Oz toward the north pole of the ecliptic. The plane of the orbit cuts the celestial sphere in the great circle NPR where N is the point where the body in its orbit crosses the ecliptic, going north; it is called the "ascending node". The angle xON (measured eastward around the ecliptic) is called the "longitude of the ascending node" and is written as Ω. The angle between the ecliptic and the plane of the orbit is the "inclination", I. For $0 \leqslant I < 90°$ the orbit is direct; for $90° < I < 180°$, it is retrograde. If OP points toward pericentron, the angle NOP = ω (measured in the sense in which the orbit is described) is called the "argument of pericentron".

These six constants are sufficient to give a geometrical picture of the orbit, and to enable position (and velocity) in the orbit to be calculated at any time. Among the alternatives often used is $\tilde{\omega} = \Omega + \omega$, called the "longitude of pericentron". (The word "pericentron" would be replaced by "perihelion", or "perigee", etc. as appropriate.)

To find the position at any time when the elements are given, first solve Kepler's equation for the appropriate value of the eccentric anomaly, and then use equations (2.1.2) to find the coordinates in the orbital reference system. The coordinates in this system can be related

to the coordinates in any other system by a series of rotations. The following successive rotations: $-\omega$ about the Z-axis, -I about the new x-axis, and $-\Omega$ about the new z-axis, will transform coordinates in the orbital reference system to those in the x-, y-, z- system of the figure. A further rotation about the x-axis through $-\epsilon$ (where ϵ is the obliquity of the ecliptic) will lead to coordinates based on the celestial equator; these are necessary if right ascension and declination are to be calculated. The transformation resulting from a rotation about an axis of reference can be most conveniently described by a matrix multiplication. For details, see Ref. 2, Appendix B. The result of all the rotations described above can be written in the form

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} P_x & Q_x \\ P_y & Q_y \\ P_z & Q_z \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}, \qquad (2.5.3)$$

where the P's and Q's are direction cosines of the X- and Y-axes with respect to the x-, y-, z-axes.

Suppose that it is required to find the geometrical elements when position and velocity, $\underline{r}_0$, $\underline{r}'_0$, are given for a time t_0. a is found from (2.4.1) and e and E_0 from (2.1.3) and (2.4.5), as before. Then T, or M_0, the mean anomaly at the epoch, can be found from Kepler's equation. The individual angles Ω, ω, and I might now be found from formulas:

$$
\begin{aligned}
\underline{h} &= \underline{r}_0 \times \underline{r}'_0 \\
&= (h_x, h_y, h_z), \\
h_x &= h\sin\Omega\sin I, \\
h_y &= -h\cos\Omega\sin I, \\
h_z &= h\cos I. \\
&f \text{ from (2.4.3)} \\
\sin u &= \frac{z}{r}\operatorname{cosec} I, \\
r\cos u &= x\cos\Omega + y\sin\Omega, \\
&(u \text{ is the "argument of the latitude"}) \\
\omega &= u - f,
\end{aligned}
\qquad (2.5.4)
$$

where an extra 360° may have to be added to make ω lie between 0 and 360°.

Alternatively, it may be better to find the P's and Q's of (2.5.3) directly. (2.5.3) can be written more generally as

$$
\begin{bmatrix} x_0 & x'_0 \\ y_0 & y'_0 \\ z_0 & z'_0 \end{bmatrix} = \begin{bmatrix} P_x & Q_x \\ P_y & Q_y \\ P_z & Q_z \end{bmatrix} \begin{bmatrix} X_0 & X'_0 \\ Y_0 & Y'_0 \end{bmatrix},
$$

where $X_0 = a(\cos E_0 - e)$, etc. from (2.1.2) and (2.4.6). Then, solving for the P's and Q's, we find (since $X_0Y'_0 - Y_0X'_0 = h$),

$$\begin{bmatrix} P_x & Q_x \\ P_y & Q_y \\ P_z & Q_z \end{bmatrix} = \begin{bmatrix} x_0 & x_0' \\ y_0 & y_0' \\ z_0 & z_0' \end{bmatrix} \begin{bmatrix} Y_0'/h & -X_0'/h \\ -Y_0/h & X_0/h \end{bmatrix} . \qquad (2.5.5)$$

The individual angles Ω, ω, and I can also be determined from the P's and Q's.

For an account of the determination of the elements when two positions for two different times are given, see Ref. 3.

In certain cases some element can only be poorly determined. For instance, if e is small, E_0 and ω or $\tilde{\omega}$ cannot be found as accurately as the other elements because, somewhere along the line, their calculation involves division by e. Similarly, if I is small, Ω is poorly determined. It is possible to put too much emphasis on the difficulties that result. ω or Ω should not be considered as goals in themselves. Suppose that the object of the work is to calculate position and velocity at any time; then it need not matter that for small e an angle such as ω is poorly determined (in fact there will be a multiplication by e during the calculation), and the accuracy of the final result need not suffer at all. Difficulties due to a small I can be avoided by using the P's and Q's. If a programmer is determined to avoid any division by e, there are several ways in which this can be achieved. One possibility is to use $e\cos E_0$ and $e\sin E_0$ as elements; there need be no doubt about their accuracy. Let E be the eccentric anomaly at time t, then from Kepler's equation applied to the

times t_0 and t, we find

$$n(t - t_0) = E - E_0 - e\sin E + e\sin E_0$$
$$= \Delta E - e\cos E_0 \sin \Delta E - e\sin E_0 \cos \Delta E + e\sin E_0, \quad (2.5.6)$$

where $\Delta E = E - E_0$. This can be solved for ΔE, and then esinE and ecosE can be calculated, and (2.5.2) and (2.5.1) used to find the position at time t.

If the elements are to be considered as slowly varying quantities in perturbed motion, other problems may arise, and different elements are needed for special cases.

Chapter Three: Expansions in Series.

3.1 Expansions in Powers of the Eccentricity.

The stumbling block in any attempt to express position in Keplerian motion explicitly in terms of the time, comes in any attempt to express the eccentric anomaly explicitly as a function of the mean anomaly. In general it cannot be done in a finite number of terms. But if the eccentricity is sufficiently small, approximate expressions can be developed that are good enough. Fortunately, nearly all the planets and satellites in the solar system have orbits with moderately small eccentricities.

For a circular orbit, $E = M$. If e is small, then, writing Kepler's equation in the form

$$E = M - e\sin E,$$

we see that to the order of e, we can put

$$E_1 = M + e\sin M.$$

Now if we put $E_2 = E_1 + \delta E_1$, and ignore e^3, we find

$$E_2 = M + e\sin M + \frac{1}{2}e^2\sin 2M.$$

Further development along these lines becomes immensely tedious, and it would be an advantage if some formula could be found that would give the general term. Such a formula is given by Lagrange's theorem,

which can be stated for the problem in hand as follows:

Let $$E = M + ef(E),$$

then $$F(E) = F(M) + \frac{e}{1!}f(M)F'(M) + \frac{e^2}{2!}\frac{d}{dM}\left\{\left[f(M)\right]^2 F'(M)\right\} + \ldots$$

$$\ldots + \frac{e^q}{q!}\frac{d^{q-1}}{dM^{q-1}}\left\{\left[f(M)\right]^q F'(M)\right\} + \ldots$$

Now put $F(E) \equiv E$, so that $F'(E) = dF/dE = 1$; and put $f(E) \equiv \sin E$. Then we get

$$E = M + \frac{e}{1!}\sin M + \frac{e^2}{2!}\frac{d}{dM}(\sin^2 M) + \ldots + \frac{e^n}{n!}\frac{d^{n-1}}{dM^{n-1}}(\sin^{n+1} M) + \ldots \quad (3.1.1)$$

Any other $F(E)$, such as $r \equiv F(E) = a(1 - e\cos E)$, can be expanded similarly.

The general statement of Lagrange's theorem would have been timely, because it includes the condition for convergence of the series (and it is not often that a question of the convergence of a series in celestial mechanics can be answered). Limitation of space prevents a discussion here, but see Ref. 4, Sec. 46. The upshot is that series in powers of the eccentricity converge for values of e less than 0.6627....

3.2 Applications of Lagrange's Theorem.

An unattractive feature of (3.1.1) is that powers of trigonometric functions appear. It is usually simpler to deal with terms such as $\sin kM$ rather than $\sin^k M$, so that Fourier series are generally preferable to power series. Also it is laborious to change from one to the

other, so that it is an advantage if a Fourier series can be generated in the first place. One way of doing this is to use the exponential function. For

$$E^{ikM} = \cos kM + i\sin kM,$$

where E is the exponential and $i^2 = -1$, so that what is generated as a power series in E^{iM} becomes a Fourier series.

Consider (2.4.3). It is usual in these developments to get rid of the square root, so we introduce

$$\frac{1+\beta}{1-\beta} = \sqrt{\frac{1+e}{1-e}} \qquad (3.2.1)$$

so that

$$\beta = \tfrac{1}{2}e(1+\beta^2). \qquad (3.2.2)$$

Also, introducing $\sin\phi = e$, we have $\beta = \tan\frac{1}{2}\phi$.

(3.2.2) could equally well have been written as

$$\beta = M + (\tfrac{1}{2}e)(1+\beta^2), \qquad M = 0.$$

Then β^j can be expanded in powers of $\frac{1}{2}e$ by Lagrange's theorem to give

$$\beta^j = \sum_{q=1}^{\infty}\left[(\tfrac{1}{2}e)^q/q!\right]\left\{\frac{d^{q-1}}{dM^{q-1}}\left[(1+M^2)^q jM^{j-1}\right]\right\}_{M=0}$$

$$= j\sum_{q=1}^{\infty}\left[(\tfrac{1}{2}e)^q/q!\right]\left\{\frac{d^{q-1}}{dM^{q-1}}\left[\sum_{p=0}^{q}\frac{q!}{(q-p)!\,p!}M^{2p+j-1}\right]\right\}_{M=0}$$

For a term to survive the operation M = 0 after the differentiation,

we must have $2p+j-1 = q-1$. Then for a definite value of p, $q = 2p+j$; so we can write

$$\beta^j = j \sum_{p=0}^{\infty} (\tfrac{1}{2}e)^{2p+j} \frac{(2p+j-1)!}{(p+j)!p!}$$

$$= (\tfrac{1}{2}e)^j \Big[1 + (\tfrac{1}{2}e)^2 j + (\tfrac{1}{2}e)^4 \frac{j(j+3)}{2!} + \ldots\Big]. \qquad (3.2.3)$$

We are now in a position to consider expansions in powers of β.

For the applications, put

$$if = \log x, \quad iE = \log y, \quad iM = \log z, \qquad (3.2.4)$$

where $i^2 = -1$, and the logs are to the base E, so that $x = E^{if}$, etc. Then

$$x^k + 1/x^k = 2\cos kf, \quad x^k - 1/x^k = 2i\sin kf, \text{ etc.}$$

From (2.4.3) and (3.2.1) we have

$$\frac{x-1}{x+1} = \frac{1+\beta}{1-\beta}\frac{y-1}{y+1},$$

so

$$x = \frac{y-\beta}{1-\beta y}, \quad \text{or} \quad y = \frac{x+\beta}{1+\beta x}. \qquad (3.2.5)$$

Then from the first of these,

$$\log x = \log y + \log(1-\beta/y) - \log(1-\beta y),$$

and, bearing in mind that for $|z|<1$, $\log(1+z) = z - z^2/2 + z^3/3 + \ldots$, we can write this as

$$\log x = \log y + \beta(y - 1/y) + \tfrac{1}{2}\beta^2(y^2 - 1/y^2) + \ldots,$$

so that, from (3.2.4) we have (after division by i)

$$f = E + 2(\beta \sin E + \tfrac{1}{2}\beta^2 \sin 2E + \tfrac{1}{3}\beta^3 \sin 3E + \ldots). \qquad (3.2.6)$$

From (3.2.5) we see that to exchange x and y it is sufficient to change the sign of β. Therefore

$$E = f - 2(\beta \sin f - \tfrac{1}{2}\beta^2 \sin 2f + \tfrac{1}{3}\beta^3 \sin 3f + \ldots). \qquad (3.2.7)$$

Substituting from (3.2.4) into Kepler's equation, we have

$$\log z = \log y - \tfrac{1}{2}e(y - 1/y).$$

Eliminating y, from (3.2.5), and using (3.2.2) to eliminate e, we can transform this to

$$\log z = \log x + \log(1 + \beta/x) - \log(1 + \beta x) - \frac{\beta}{1+\beta^2}\,\frac{(1+\beta^2)(x^2-1)}{(x+\beta)(1+\beta x)}.$$

Now $\frac{1-\beta^2}{1+\beta^2} = \sqrt{1 - e^2} = \cos\phi$, so that the final term on the right hand side can be written as

$$\beta\cos\phi\left[x/(1+\beta x) - (1/x)/(1+\beta/x)\right]$$

$$= \beta\cos\phi\left[\tfrac{1}{x}(1 - \beta/x + \beta^2/x^2 - \ldots) - x(1 - \beta x + \beta^2 x^2 - \ldots)\right].$$

Therefore, expanding the logarithms as before, and substituting from (3.2.4), we get

$$M = f - 2(\beta\sin f - \tfrac{1}{2}\beta^2\sin 2f + \ldots) + 2(\beta\cos\phi(-\sin f + \beta\sin 2f - \ldots)$$

$$= f - 2\left[\beta(1+\cos\phi)\sin f - \beta^2(\tfrac{1}{2}+\cos\phi)\sin 2f + \beta^3(\tfrac{1}{3}+\cos\phi)\sin 3f + \ldots\right]. \tag{3.2.8}$$

The difference between the mean and true anomalies is called the "equation of the center".

3.3 Fourier Series.

The derivation of the series in the preceding section was a trifle roundabout. Before proceeding with the direct derivation of Fourier series we shall briefly state enough theorems to build up the relevant background.

Let $f(t)$ be a periodic function with bounded variation and period 2π; let it be integrable for all t, so that the products $f(t)\sin pt$, $f(t)\cos pt$ are also integrable. Define

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(t)\,dt,$$

and

$$a_p = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos pt\,dt, \qquad b_p = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin pt\,dt.$$

The series

$$a_0 + \sum_{p=1}^{\infty}(a_p\cos pt + b_p\sin pt) \tag{3.3.1}$$

is called the Fourier series of $f(t)$. If $f(t)$ is continuous, its sum is equal to $f(t)$. Furthermore, if its derivative is bounded, then the Fourier series is uniformly convergent.

If the form (3.3.1) is accepted, then the formula for the coefficients is very easily recovered by multiplying through by cospt or sinpt and integrating from 0 to 2π, so that every term but one vanishes. If f(t) is an even function, then only the a_p appear, and it is sufficient to integrate from 0 to π, and divide by $\pi/2$. Similarly, if f(t) is an odd function, only the b_p appear.

Using the exponential function, we could also put

$$f(t) = \sum_{p=-\infty}^{+\infty} \alpha_p E^{ipt},$$

where

$$\alpha_p = \frac{1}{2\pi} \int_0^{2\pi} E^{ipt} f(t)\, dt. \qquad (3.3.2)$$

(Note that as soon as trigonometric functions are replaced by exponential functions, the summations must go from minus infinity to plus infinity.)

Consider the expansion of the function a/r as a Fourier series in the mean anomaly. It is an even function of E, and consequently of M. Also a/r = dE/dM. Therefore

$$\frac{a}{r} = \frac{1}{\pi} \int_0^{\pi} \frac{a}{r} dM + \frac{2}{\pi} \sum_{p=1}^{\infty} \cos pM \int_0^{\pi} \frac{a}{r} \cos pM\, dM$$

$$= \frac{1}{\pi} \int_0^{\pi} dE + \frac{2}{\pi} \sum_{p=1}^{\infty} \cos pM \int_0^{\pi} \cos(pE - pe\sin E) dE.$$

Define the "Bessel's coefficient" $J_p(x)$, or order p and argument x by

$$J_p(x) = \frac{1}{\pi}\int_0^{\pi} \cos(p\phi - x\sin\phi)\, d\phi. \qquad (3.3.3)$$

Then we can write

$$\frac{a}{r} = 1 + 2\sum_{p=1}^{\infty} J_p(pe)\cos pM. \qquad (3.3.4)$$

These coefficients are ubiquitous, and it is necessary to break off and derive some of their properties before continuing to develop any other series.

3.4 Properties of Bessel's Functions.

$J_p(x)$ was defined in (3.3.3). But since

$$\frac{1}{2\pi}\int_0^{2\pi} \sin(p\phi - x\sin\phi)\, d\phi = 0,$$

we could have written

$$J_p(x) = \frac{1}{2\pi}\int_0^{2\pi} E^{-ip\phi}E^{ix\sin\phi}d\phi. \qquad (3.4.1)$$

Now suppose that we wanted to expand the function $E^{ix\sin\phi}$ as

$$E^{ix\sin\phi} = \sum_{p=-\infty}^{+\infty} \alpha_p E^{ip\phi}.$$

Then, from (3.3.2) we would find that $\alpha_p = J_p$, so that

$$E^{ix\sin\phi} = \sum_{-\infty}^{+\infty} J_p(x)E^{ip\phi}, \qquad (3.4.2)$$

a formula that can be useful, incidentally, where trigonometric functions of trigonometric functions are concerned.

Now put $E^{i\phi} = z$, so that $2i\sin\phi = z - 1/z$. Then (3.4.2) becomes

$$\exp\left[\frac{x}{2}(z - 1/z)\right] = \sum_{-\infty}^{+\infty} J_p(x)z^p. \qquad (3.4.3)$$

The left hand side of (3.4.3) can be written as the product of

$$\sum_{\alpha=0}^{\infty}(\tfrac{1}{2}x)^\alpha z^\alpha/\alpha! \quad \text{and} \quad \sum_{\beta=0}^{\infty}(-1)^\beta(\tfrac{1}{2}x)^\beta z^{-\beta}/\beta!.$$

To find the coefficient of z^p put $\alpha = \beta + p$. Now α cannot be negative, so that for $p \geqslant 0$,

$$J_p(x) = \sum_{\beta=0}^{\infty}(-1)^\beta \frac{1}{\beta!(\beta + p)!}(\tfrac{1}{2}x)^{p+2\beta}. \qquad (3.4.4)$$

For $p < 0$, the summation runs from $\beta = -p, -p+1, \ldots$. The series (3.4.4) is absolutely convergent for all x.

In (3.4.3) change z to -z, and x to -x; the left hand side is the same, so that

$$J_p(x) = (-1)^p J_p(-x).$$

Also, change z to -1/z. The left hand side is still the same, so that

$$J_p(x) = (-1)^p J_{-p}(x).$$

Combining these two results, we find

$$J_{-p}(-x) = J_p(x). \qquad (3.4.5)$$

Differentiating (3.4.3) with respect to z, and using (3.4.3) to

remove the exponential on the left hand side, we get

$$\tfrac{1}{2}x(1 + 1/z^2) \sum J_p(x)\, z^p = \sum pJ_p(x)\, z^{p-1}.$$

So that from the coefficients of z^{p-1}, we find

$$\tfrac{1}{2}x\left[J_{p-1}(x) + J_{p+1}(x)\right] = pJ_p(x). \qquad (3.4.6)$$

Similarly, differentiating (3.4.3) with respect to x, and considering the coefficients of z^p, we can find

$$\tfrac{1}{2}\left[J_{p-1}(x) - J_{p+1}(x)\right] = J_p'(x). \qquad (3.4.7)$$

Differentiating (3.4.7) with respect to x, we have

$$\begin{aligned} J_p''(x) &= \tfrac{1}{2}\left[J_{p-1}'(x) - J_{p+1}'(x)\right] \\ &= \tfrac{1}{4}\left[J_{p-2}(x) - 2J_p(x) + J_{p+2}(x)\right] && \left[\text{from (3.4.7)}\right] \\ &= \tfrac{1}{4}\left[\tfrac{2}{x}(p-1)J_{p-1}(x) - J_p(x) - 2J_p(x) + \tfrac{2}{x}(p+1)J_{p+1}(x) - J_p(x)\right] \\ &\qquad && \left[\text{from (3.4.6)}\right] \\ &= -J_p(x) + \tfrac{1}{2x}\left[(p-1)J_{p-1}(x) + (p+1)J_{p+1}(x)\right] \\ &= -J_p(x) + \frac{p^2}{x^2} J_p(x) - \tfrac{1}{x} J_p'(x). && \left[\text{from (3.4.7)}\right] \end{aligned}$$

So J_p is a solution of the equation

$$y'' + \tfrac{1}{x}y' + (1 - p^2/x^2)y = 0.$$

The general theory of Bessel's functions can start from this equation; but this is not needed for our purpose. We need only the solutions of the first kind, with integral values of p, and the definition given

above is sufficient.

The series (3.4.4) demonstrates that the J_p can always be calculated. But there are many alternative methods of calculation, using such devices as recurrence relations, or continued fractions. See Ref. 1.

3.5 Applications of Bessel's Functions.

Consider the expansion of sinmE. It is an odd function of E or M, so that

$$\text{sinmE} = \frac{2}{\pi} \sum_{p=1}^{\infty} \text{sinpM} \int_0^{\pi} \text{sinmE sinpM dM}.$$

Now $\text{sinpM dM} = -\frac{1}{p}\text{d(cospM)}$, so that, introducing E, we can write

$$\text{sinmE} = -\frac{2}{\pi} \sum \frac{\text{sinpM}}{p} \int_0^{\pi} \text{sinmE d}\Big[\cos(\text{pE} - \text{pesinE})\Big],$$

and, integrating by parts,

$$\text{sinmE} = -\frac{2}{\pi} \sum \frac{\text{sinpM}}{p} \left\{ \Big[\text{sinmE cos}(\text{pE} - \text{pesinE})\Big]_0^{\pi} - \int_0^{\pi} \text{mcosmE cos}(\text{pE} - \text{pesinE})\ \text{dE} \right\}.$$

The integrated term vanishes at the limits; using the formula for the product of two cosines, the integrand can be developed to give

$$\begin{aligned}\text{sinmE} &= \frac{m}{\pi} \sum \frac{\text{sinpM}}{p} \int_0^{\pi} \left\{ \cos\Big[(p+m)E - \text{pesinE}\Big] + \cos\Big[(p-m)E - \text{pesinE}\Big] \right\} \text{dE} \\ &= m \sum_{p=1}^{\infty} \frac{\text{sinpM}}{p} \left\{ J_{p-m}(\text{pe}) + J_{p+m}(\text{pe}) \right\}.\end{aligned}$$

When m = 1, we have, by (3.4.6),

$$\text{sinE} = \frac{2}{e} \sum_{p=1}^{\infty} \frac{\text{sinpM}}{p} J_p(\text{pe}). \qquad (3.5.1)$$

Similarly, we find

$$\cos mE = a_0 + \frac{2}{\pi}\sum_{p=1}^{\infty} \cos pM \int_0^{\pi} \cos mE \cos pM \, dM$$

$$= a_0 + \frac{2}{\pi}\sum \cos pM \int_0^{\pi} \frac{m}{p} \sin mE \sin(pE - pe\sin E)\, dE$$

(after integration by parts)

$$= a_0 + m\sum \frac{\cos pM}{p}\left\{ J_{p-m}(pe) - J_{p+m}(pe)\right\}.$$

Here

$$a_0 = \frac{1}{\pi}\int_0^{\pi} \cos mE \, dM = \frac{1}{\pi}\int_0^{\pi} \cos mE\,(1 - e\cos E)\, dE$$

$$= \frac{1}{\pi}\int_0^{\pi} \left[\cos mE - \frac{1}{2}e\cos(m+1)E - \frac{1}{2}e\cos(m-1)E\right] dE$$

$$= 1 \text{ if } m = 0;\ -e/2 \text{ if } m = 1;\ 0 \text{ if } m > 1.$$

In particular, using (3.4.7),

$$\cos E = -\frac{1}{2}e + 2\sum \frac{\cos pM}{p} J_p'(pe). \qquad (3.5.2)$$

We now have enough formulas to expand quite a lot of functions as Fourier series in the mean anomaly. For instance, Kepler's equation combined with (3.5.1) will cope with E. r/a can be expanded using (2.1.3) and (3.5.2). X and Y of (2.1.2) can be found similarly. Sometimes a little ingenuity can help; in seeing, for instance, that $X/r^3 = -a^{-3}d^2x/dM^2$, $Y/r^3 = -a^{-3}d^2Y/dM^2$. Another example is

$$\sin f = \sqrt{1-e^2}\,\sin E/(1 - e\cos E) = \cot\phi\, \frac{d}{dM}\left(\frac{r}{a}\right).$$

cosf is easily found from (2.1.1) and (3.3.4). A function such as $(r/a)^2$ can be easily written down in terms of a Fourier series in E, and from there to one in M. And so on. Many more examples are given

in Refs. 1 and 4.

It should be noted that these Fourier series are valid for any value of the eccentricity; but if they are re-arranged as power series in the eccentricity, then the upper limit noted in Sec. 3.1 applies.

In the series for a/r or r/a or powers of these, it is noticeable that the lowest power of e in any coefficient is equal to the multiple of M in that term; this fact is a great help when deciding where to truncate a series. Although the equality just pointed out is not general, the fact that the lowest order of e increases as the coefficient of M is; this is a characteristic of these expansions stressed by D'Alembert, which now bears his name. In the expansion of $(a/r)^{\pm k}$ times the sine or cosine of m times the eccentric or true anomaly, the lowest power of e is in general equal to the coefficient of M minus m.

Postscript.

My impression on reading these notes is that they are parlously incomplete. No mention has been made of expansions in powers of the time; nor of the first-order differences between two "nearly equal" elliptic orbits. But these fall usually under the heading of "orbit determination" and are dealt with more than adequately in Ref. 3. Nothing has been said about the proper choice of units, even though, without this, an attempt at practical calculation in celestial mechanics may be stillborn. Also hyperbolic orbits have been neglected, in spite of their increasing importance. Bearing in mind these and other omissions, the reader should redress the balance by consulting some of the references.

References.

(1) Brouwer, D., and Clemence, G.M. "Methods of Celestial Mechanics". New York: Academic Press, 1961.

(2) Danby, J.M.A. "Fundamentals of Celestial Mechanics". New York: Macmillan, 1962.

(3) Herget, P. "The Computation of Orbits". University of Cincinnati, 1948.

(4) Plummer, H.C. "Dynamical Astronomy". New York: Dover Publications, 1960. (First published, 1918.)

(5) Clemence, G.M. "Planetary Distances According to General Relativity". Astron. J., 67, 379, 1962.

by J.M.A. Danby

Consider the system of n first order differential equations

$$dX_i/dt = f_i(X_1, X_2, \ldots, X_n; t), \quad i = 1, 2, \ldots n, \qquad (1)$$

relating the n coordinates X_i and the time t. These can be written symbolically in the condensed form

$$\underline{X}' = \underline{f}(\underline{X},t), \qquad (1A)$$

where $\underline{X}$ and $\underline{f}$ are column matrices; the primes represent differentiation with respect to the time.

Suppose that a solution $\underline{X}_R(t)$ has been found, having initial conditions $\underline{X}_R(t_0) = \underline{X}_0$. A "slightly different" solution, $\underline{X}_R + \delta\underline{X}$, might be found by solving equations (1) again, subject to initial conditions $\underline{X}_0 + \delta\underline{X}_0$ at t_0. Then $\delta\underline{X}$ would be found by subtracting $\underline{X}_R$. But this approach can be extravagant in significant figures, and it is often better to solve directly for $\delta\underline{X}$.

If the squares and products of small quantities are neglected, then $\delta\underline{X}$ must satisfy the first variational equations of the system (1):

$$\begin{bmatrix} \delta X_1' \\ \delta X_2' \\ \cdot \\ \cdot \\ \cdot \\ \delta X_n' \end{bmatrix} = \begin{bmatrix} \partial f_1/\partial X_1 & \partial f_1/\partial X_2 & \cdots & \partial f_1/\partial X_n \\ \partial f_2/\partial X_1 & \partial f_2/\partial X_2 & \cdots & \partial f_2/\partial X_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \partial f_n/\partial X_1 & \partial f_n/\partial X_2 & \cdots & \partial f_n/\partial X_n \end{bmatrix} \begin{bmatrix} \delta X_1 \\ \delta X_2 \\ \cdot \\ \cdot \\ \cdot \\ \delta X_n \end{bmatrix} \tag{2}$$

or

$$\delta \underline{X}' = A \delta \underline{X}. \tag{2A}$$

The solution $\underline{X}_R(t)$ will be called the "reference orbit". Each of the partial differential coefficients in the n-by-n matrix A is evaluated along the reference orbit, so that A is a known function of the time.

Equations (2) are solved when any set of n linearly independent solutions is known. Finding these may present difficulties; but suppose for the moment that we have such a set, and that it consists of the separate columns of the matrix with elements $\delta x_{ij}(t)$. Since any linear combination of these columns also gives a solution, the columns of

$$\Omega(t_0,t) \equiv \left[\delta x_{ij}(t)\right] \left[\delta x_{ij}(t_0)\right]^{-1} \tag{3}$$

must all be solutions. The matrix $\Omega(t_0,t)$ is equal to the identity matrix when $t = t_0$; this provides the necessary initial conditions to find its components by numerical integration. For example equations (2) would be solved subject to the initial conditions $\delta X_1 = 1$,

$\delta X_i = 0$ $(i \neq 1)$ to give the first column. The initial conditions for the second column would be $\delta X_2 = 1, \delta X_i = 0$ $(i \neq 2)$; and so on. $\Omega(t_0,t)$ is called the "matrizant" (or "fundamental solution matrix" or "state transition matrix") of the system (2). Since each of its columns satisfies (2), it must itself satisfy

$$\Omega' = A\Omega, \tag{4}$$

where $$\Omega(t_0,t_0) = I.$$

If the function $\delta\underline{X}(t)$ were to have initial conditions $\delta\underline{X}(t_0) = \delta\underline{X}_0$, then the appropriate solution of (2) would be

$$\delta\underline{X}(t) = \Omega(t_0,t)\delta\underline{X}_0. \tag{5}$$

It is clear that $\Omega(t_0,t)$ is the Jacobian matrix with components $\partial X_i/\partial X_{0,j}$, etc.

(Matrizants were introduced by Peano and Baker. Their theory is discussed in the Summer Institute Notes of 1960, p. 95, et seq., and in many texts on differential equations; but this theory is not directly relevant to the present discussion.)

Consider the relations between residuals $\delta\underline{X}$ at times t_0, t_1, and t_2. We have

$$\delta\underline{X}_2 = \Omega(t_0,t_2)\delta\underline{X}_0,$$

and

$$\delta\underline{X}_2 = \Omega(t_1,t_2)\delta\underline{X}_1$$
$$= \Omega(t_1,t_2)\Omega(t_0,t_1)\delta\underline{X}_0.$$

Therefore

$$\Omega(t_0,t_2) = \Omega(t_1,t_2)\Omega(t_0,t_1), \tag{6}$$

a result that is also evident from the fact that $\Omega(t_0,t)$ is a Jacobian matrix.

Consider the application of the matrizant to some situations in the context of astronautics. Suppose that a reference orbit has been calculated. If some maximum permissible error at time t_1 is specified, then the maximum permissible error at any earlier time t_0 can be calculated if $\Omega(t_1,t_0)$ is known. If an error is observed at t_0, the effect at a later time t_1 can be calculated using $\Omega(t_0,t_1)$. But if t_1 is fixed and t_0 varies it is obviously inconvenient to solve the equations for $\Omega(t_0,t_1)$ many times for different t_0, and it is better to put

$$\delta \underset{\sim}{x}_1 = \Omega^{-1}(t_1,t_0)\delta \underset{\sim}{x}_0, \tag{7}$$

and solve the corresponding equations with the initial conditions applied at t_1. We notice, incidentally, that $\Omega^{-1}(t_1,t_0) = \Omega(t_0,t_1)$. Furthermore, it is possible to avoid the inversion of the matrix; for let

$$\gamma(t_1,t)\Omega(t_1,t) = I.$$

Differentiating with respect to t, and using (4), we find

$$\gamma' = -\gamma A. \tag{8}$$

Equation (8) is called the "adjoint equation" of (4). (The use of the adjoint equations in this sort of context was first cultivated in

ballistics, and is described in "Mathematics for Exterior Ballistics" by G.A. Bliss, Wiley, 1944.)

Now suppose that equations (4) and (8) have both been solved, the initial conditions making each matrix equal to the identity matrix at time t_1. Writing (6) as

$$\begin{aligned}\Omega(t_0,t_2) &= \Omega(t_1,t_2)\Omega^{-1}(t_1,t_0) \\ &= \Omega(t_1,t_2)\gamma(t_1,t_0),\end{aligned} \tag{6A}$$

we see that the matrizant relating <u>any</u> two times can be found.

Normally $\underline{X}$ will have six components, of position

$$\underline{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \text{ and velocity } \underline{r}' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} X_4 \\ X_5 \\ X_6 \end{bmatrix}.$$

Let the matrizant in (5) be subdivided into four three-by-three matrices:

$$\Omega(t_0,t) = \begin{bmatrix} U(t_0,t) & V(t_0,t) \\ W(t_0,t) & Y(t_0,t) \end{bmatrix}. \tag{9}$$

Suppose that an error $\delta\underline{r}_0$ is observed at t_0, and it is required that <u>after</u> a thrust has been applied there will be a velocity residual $\delta\underline{r}_0'$ such that $\delta\underline{r}$ at time t is zero. Then we have

$$\delta\underline{r}_0' = -V^{-1}U\delta\underline{r}_0. \tag{10}$$

Consider motion subject to a force function R. The differential

equations of motion are

$$X_1' = X_4,\ X_2' = X_5,\ X_3' = X_6,\ X_4' = \partial R/\partial X_1,\ X_5' = \partial R/\partial X_2,\ X_6' = \partial R/\partial X_3.$$

The first variational equations can be written as

$$\delta \underline{X}' = A\delta \underline{X},$$

where

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \partial^2 R/\partial X_1^2 & \partial^2 R/\partial X_1 \partial X_2 & \partial^2 R/\partial X_1 \partial X_3 & 0 & 0 & 0 \\ \partial^2 R/\partial X_1 \partial X_2 & \partial^2 R/\partial X_2^2 & \partial^2 R/\partial X_2 \partial X_3 & 0 & 0 & 0 \\ \partial^2 R/\partial X_3 \partial X_1 & \partial^2 R/\partial X_3 \partial X_2 & \partial^2 R/\partial X_3^2 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & I \\ B & 0 \end{bmatrix}, \text{ say.} \qquad (11)$$

Then substituting from (9) and (11) into the equation $\Omega' = A\Omega$, we find

$$U' = W,\quad V' = Y,\quad W' = BU,\quad Y' = BV,$$

from which

$$U'' = BU \quad \text{and} \quad V'' = BV. \qquad (12)$$

Equations (12) are to be solved subject to the initial conditions

$$\left.\begin{aligned} U(t_0,t_0) &= I, \quad V(t_0,t_0) = 0; \\ U'(t_0,t_0) &= 0, \quad V'(t_0,t_0) = I. \end{aligned}\right] \qquad (13)$$

The columns of U and V are six linearly independent solutions of the equation

$$\delta \underline{r}'' = B\delta\underline{r}, \tag{14}$$

which is the first variational equation of the equation of motion in the form $\underline{r}'' = \text{grad}R$.

Let $\delta \underline{r}_0$ and $\delta \underline{r}_0'$ be the initial increments in position and velocity to be applied to the reference orbit at time t_0. Then at time t

$$\delta \underline{r} = U(t_0,t)\delta\underline{r}_0 + V(t_0,t)\delta\underline{r}_0'. \tag{15}$$

$\delta \underline{r}$ is a solution of (14), and since the components of $\delta \underline{r}_0$ and $\delta \underline{r}_0'$ can be considered as six independent, arbitrary constants, it is clear that (15) is the general solution of (14).

The matrizant and its components should always be considered as functions of two variables (two independent times). Now consider

$$Z(t_0,t) = \int_{t_0}^{t} U(\tau,t)\,d\tau.$$

We have

$$\partial Z/\partial t = I + \int_{t_0}^{t} \left|\partial U(\tau,t)/\partial t\right| d\tau$$

and

$$\partial^2 Z/\partial t^2 = \int_{t_0}^{t} \left|\partial^2 U(\tau,t)/\partial t^2\right| d\tau$$

$$= \int_{t_0}^{t} B(t)U(\tau,t)\,d\tau$$

$$= BZ.$$

So Z satisfies the differential equation as well as the initial conditions for V, and must therefore be identical with V. Hence

$$U(t_0,t) = -\partial V(t_0,t)/\partial t_0. \tag{16}$$

So the matrizant (9) can be written

$$\Omega(t_0,t) = \begin{bmatrix} -\partial V/\partial t_0 & V(t_0,t) \\ -\partial^2 V/\partial t \partial t_0 & \partial V/\partial t \end{bmatrix} \tag{17}$$

Let $\Omega^{-1}(t_0,t) = \gamma(t_0,t) = \begin{bmatrix} \bar{U} & \bar{V} \\ \bar{W} & \bar{Y} \end{bmatrix}$.

Then from (8) and (11) we find

$$\begin{bmatrix} \bar{U}' & \bar{V}' \\ \bar{W}' & \bar{Y}' \end{bmatrix} = -\begin{bmatrix} \bar{U} & \bar{V} \\ \bar{W} & \bar{Y} \end{bmatrix} \begin{bmatrix} 0 & I \\ B & 0 \end{bmatrix}.$$

Therefore

$$\bar{V}' = -\bar{U}, \quad \bar{Y}' = -\bar{W}, \quad \bar{U}' = -\bar{V}B, \quad \bar{W}' = -\bar{Y}B.$$

So

$$\bar{V}'' = \bar{V}B, \quad \bar{Y}'' = \bar{Y}B, \tag{18}$$

where

$$\left.\begin{aligned} \bar{V}(t_0,t_0) &= 0, & \bar{Y}(t_0,t_0) &= I, \\ \bar{V}'(t_0,t_0) &= -I, & \bar{Y}'(t_0,t_0) &= 0. \end{aligned}\right\} \tag{19}$$

Now B is symmetrical, so that transposing equations (18), we find

$$\bar{V}''^{T} = B\bar{V}^{T}, \quad \bar{Y}''^{T} = B\bar{Y}^{T}. \tag{18A}$$

Comparing (18A) and (19) with (12) and (13), we see that

$$\bar{V} = -V^T, \quad \bar{U} = Y^T, \quad \bar{W} = -W^T, \quad \text{and} \quad \bar{Y} = U^T.$$

Therefore

$$\Omega^{-1}(t_0,t) = \begin{bmatrix} U & V \\ W & Y \end{bmatrix}^{-1} = \begin{bmatrix} Y^T & -V^T \\ -W^T & U^T \end{bmatrix}, \tag{20}$$

a result that applies only to suitable equations arising from motion in the conservative field of force.

The components of a matrizant would normally have to be found numerically; but in some cases it is possible to find them analytically. This is notably so in the case of Keplerian motion, for which the components of V are given in a paper in A.J. 67, June, 1962. This matrizant has possible applications in perturbation problems in celestial mechanics. The components are most easily found, not by solving the differential equation, but by considering, from first principles, what the effects of errors in velocity at time t_0 will be on errors in position at time t.

Consider the equation

$$\delta \underline{X}' = A\delta\underline{X} + \underline{g}(t), \tag{21}$$

in which a "forcing function", $\underline{g}(t)$, has been added to (2). The equation is no longer homogeneous, and one way to solve it is to take the solution of the homogeneous part, viz.

$$\delta\underline{X}(t) = \Omega(t_0,t)\delta\underline{X}_0, \tag{22}$$

and allow the arbitrary constants, $\delta \underline{X}_0$, to vary. Then

$$\delta \underline{X}' = \Omega' \delta \underline{X}_0 + \Omega \delta \underline{X}_0'$$
$$= A\Omega \delta \underline{X}_0 + \Omega \delta \underline{X}_0'.$$

Substituting this, and (5), into (21), we have

$$A\Omega \delta \underline{X}_0 + \Omega \delta \underline{X}_0' = A\Omega \delta \underline{X}_0 + \underline{g},$$

so that

$$\delta \underline{X}_0 = \int_{t_0}^{t} \Omega^{-1} \underline{g} dt.$$

The complete and general solution is therefore

$$\delta \underline{X} = \Omega(t_0,t)\delta \underline{X}_0 + \Omega(t_0,t)\int_{t_0}^{t} \Omega^{-1}(t_0,\tau)\underline{g}(\tau)d\tau, \qquad (23)$$

where $\delta \underline{X}_0$ is once again constant. This is the exact solution of (21), subject to the initial conditions $\delta \underline{X}(t_0) = \delta \underline{X}_0$; no conditions about orders of magnitude are imposed. The first term, which includes the arbitrary constants, is the complementary function, and the second is the particular integral. If the particular integral is to be found numerically, probably the best procedure is to solve equations (21) subject to the initial conditions $\delta \underline{X}(t_0) = 0$. (23) can be simplified by the use of the multiplication formula (6) to give

$$\delta \underline{X} = \Omega(t_0,t)\delta \underline{X}_0 + \int_{t_0}^{t} \Omega(\tau,t)\underline{g}(\tau)d\tau. \qquad (24)$$

Another form is

$$\delta \underline{X} = \Omega(t_0,t)\delta \underline{X}_0 + \Omega(s,t)\int_{t_0}^{t} \Omega(\tau,s)\underline{g}(\tau)d\tau, \qquad (25)$$

where s is an arbitrary time, chosen to make Ω as simple as possible.

If the equations of motion are in cartesian coordinates, then the first three components of $\underline{g}$ are zero; writing the last three as $\underline{f}$, we have, from (24)

$$\delta\underline{r} = U(t_0,t)\delta\underline{r}_0 + V(t_0,t)\delta\underline{r}_0' + \int_{t_0}^{t} V(\tau,t)\underline{f}(\tau)d\tau. \qquad (26)$$

Or, from (25) and (20),

$$\delta\underline{r} = U(t_0,t)\delta\underline{r}_0 + V(t_0,t)\delta\underline{r}_0' + \begin{bmatrix} U(s,t) & V(s,t) \end{bmatrix} \int_{t_0}^{t} \begin{bmatrix} -V^T(s,\tau) \\ U^T(s,\tau) \end{bmatrix} \underline{f}(\tau)d\tau. \qquad (27)$$

In the case of disturbed Keplerian motion, s would certainly be a time of perihelion passage. Also in this case there are advantages in changing the independent variable from the time to the eccentric anomaly in the reference orbit.